KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Bioscience Engineering

A framework for the creation and exploration of cross platform expression compendia

Qiang Fu

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Bioscience Engineering

14th October 2014

A framework for the creation and exploration of cross platform expression compendia

Qiang FU

Examination committee:
Prof. dr. ir. Johan Buyse, chair
Prof. dr. ir. Kathleen Marchal, supervisor
Prof. dr. ir. Kristof Engelen, co-supervisor
(Fondazione Edmund Mach, Italy)
Prof. dr. ir. Dirk Springael
Prof. dr. ir. Yves Moreau
Prof. dr. ir. Jozef Vanderleyden
Dr. ir. Maarten Fauvart

Prof. dr. Klaas Vandepoele
(Ghent University, Belgium)

Dissertation presented in partial
fulfillment of the requirements for
the degree of Doctor
in Bioscience Engineering

14th October 2014

ISBN 978-90-8826-381-1

D/2014/11109/55

Preface

Many years ago, when I was looking for a new challenge, a little advertisement printed in A4 in the corridor caught my attention: 'Software engineer technician for bioinformatics'. I loved biology in high school and it is the science of 21th century. This is a great opportunity to extend my career into this exciting domain. I sent my CV without hesitation, and luckily, I was hired. Later, encouraged by Prof. Kathleen Marchal and Dr. Kristof Engelen, I started pursuing a PhD degree. It has been an exciting, however, challenging, difficult, and stressful journey. Now, it finally reached the happy ending. I would like to use this opportunity to acknowledge all the people whose support, understanding, encouragement, and company has helped me out of all the tough moments, and provided me with an enjoyable life.

My deepest gratitude goes to my promoter Prof. Kathleen Marchal and co-promotor Dr. Kristof Engelen. Thanks them to give me a great opportunity to pursue my PhD in this exciting field of bioinformatics. Kristof has closely supervised my project from day one. The countless discussions we had together has greatly helped me in expanding my scientific knowhow, and in focusing on the right path to progress my research. Kathleen, although not involved in daily supervise, has provided critical suggestions through out my PhD study. Her meticulous attitude towards science and hard work spirit has been the model that guides me through a difficult and sometimes boring research life. My success would not be possible without their help and guidance.

I do wish to express my sincere gratitude to the accessors of my research Prof. Dirk Springael and Yves Moreau for their kindly support and valuable feedbacks at different stages of my research, and their suggestions on the thesis. I would also like to thank Prof. Jozef Vanderleyden, Klaas Vandepoele, and Dr. Maarten Fauvart for their willingness to invest their precious time as my jury member. Their comments has greatly helped improve the quality of the thesis. Special thanks goes to Chairman Prof. Ivo Vankelecom and Prof. Johan Buyse.

I would like to acknowledge those colleagues who I have worked closely with during my PhD study. Carolina gave many guidance in every projects we worked together, and helped me resolve many difficult issues. Pieter is smart and energetic. He has contributed greatly in developing various software systems. Aminael, the biology specialist in the team, always provided answer for any questions I throw to him precisely and swiftly. He has helped me gain many insights into the biology behind the data. My sincere thanks also goes to all the former and current members of bioinformatics group, Abeer, Alejandra, Bram, Daniel, Dries, Sun Hong, Zhao Hui, Inge, Ivan, Karen, Lore, Lucia, Lyn, Marleen, Nicolas, Paymen, Pieter Monsieurs, Riet, Segio, Thomas, Tim, Valeria. I will surely miss the conferences, brainstorms, and midday lunches where, together, we discussed all kinds of interesting topics, played many fun games, and made many stories.

I would also like to take this opportunity to thank all my friends who greatly enriched my social life in Leuven. Some of them are my international friends including Keon, Maria, Leen, Nicolas, Paula, Roxana, Tudor, Sandra, Siddi, etc. And many more from the Chinese community, Yang Qu, Minli Li, Yanru Shi, Caiping Li, Zhiyu Chen, Jing Su, Yinli Kan, Beiwen Chen, Ying He, Jianxiong Sheng, Ping Hou, Ye Guan, Ye Su, Liu Fang, Meixun Jin, Geng Chen, Andre Davis, Zhiyong Zhang, Yaqin Chen, and many others. Special thanks are given to Christophe Morren and Hervé van der aerschot. They introduced me to two fantastic sports, funky jazz dance and badminton, which, ever since, became an indispensable part of my life, and helped me to overcome the ever mounting stress. Oh, yes, last but not least, Yves van de peer, the BOSS of boss, who always encouraged me to challenge ever harder piste, and many friends I made though those exciting ski trips. Thank you very much for all your support, help and company. It has made the past years of my life such a wonderful experience.

Last, but certainly the most, I want to thank my family. My father and mother have been always supportive and caring throughout my life, and ever more from afar after I came to Belgium. My parents-in-law, who traveled so far away from home for the first time in their life to Belgium, helped us out during the busiest moment right before and after Vivien was born. My wife, Ruixia, always understands me and tolerates my tempers. Her unconditional love and support has helped me get through many difficult times. My lovely little daughter, Vivien, brings lots of joy and happiness everyday. Words can not express my gratitude to my family for all your love and encouragement. I dedicate this work to you.

Qiang,
October, 2014

Abstract

With the unprecedented ability to systematically probe gene expression at the genome scale, microarrays have become an indispensable technology adopted by most of the laboratories across the world, generating a wealth of data for a variety of species. Although comprehensive in the gene dimension, any microarray based study alone provides a limited scope at the level of condition. However combining expression data from different labs provides the opportunity to investigate gene expression of a particular species at a more global level, and to view a specific study from the perspective of existing knowledge. The goal of this research is developing a novel methodology and system to explore this opportunity.

We first developed a methodology to create an organism-specific cross-platform compendium based on publicly available gene expression data. Special attention has been paid to facilitate automated data retrieval by resolving heterogeneities in the data representation, and to improve data consistency and compatibility through systematic renormalization of the data. Compared with existing single platform compendia, our methodology provides a broader range of data due to its cross platform nature.

Using this novel methodology, we constructed three comprehensive expression compendia for the bacterial model organisms, *Escherichia coli*, *Bacillus subtilis*, and *Salmonella enterica* serovar Typhimurium. Moreover, efforts have been taken to create a web portal with intuitive functionalities for data analysis and visualization, providing public access to these three compendia.

One of the most important applications of compendia is to study the response of an organism to environmental changes by identifying condition dependent functional modules and studying the underlying regulatory mechanisms responsible for the observed expression variations. Different methods exist for this purpose. Each makes distinct assumptions to handle the under-deterministic nature of this complex problem, and consequently generates complementary

results. Here, we demonstrated such complementarity between two methods, DISTILLER and COLOMBOS, in a case study, in which co-expression modules containing gene *sodA* are extracted from the *E. coli* compendium using each method, and compared against each other. Through this example, we stress the importance of choosing the right method based on the research purpose.

At last, we extended the methodology to handle the increased complexity of the monocot *Zea mays*, specifically addressing the following two issues: resolving inconsistency in the platform-probe annotation, and providing a more precise biological sample annotation, which can reflect the different genetic repositories of maize (breeding lines), the complexity of plant's life style (development stage), and its more complex tissue structure. We further upgraded the web access portal accordingly with new functions adapted to queries specific for a higher organism like *Zea mays*.

Beknopte samenvatting

Het meten van genexpressie op een genoom schaal is ondertussen routine is geworden in de meeste moleculaire laboratoria. Hierdoor is in het publiek domein een onschatbare hoeveelheid genexpressie metingen beschikbaar voor diverse modelorganismen. Hoewel uitgebreid in de gendimensie, beperkt elke individuele genexpressie studie zich tot het meten van de expressie in een enkele conditie. Door data van verschillende labo's te combineren in een enkel expressie compendium kan een meer globaal beeld bekomen worden van de conditie afhankelijkheid van transcriptionele regulatie en dus genexpressie.

De bedoeling van deze thesis was dan ook om een methodologie te ontwikkelen die het toelaat om aan de hand van publiek beschikbare expressie gegevens op een semi-automatische wijze een organisme-specifiek expressiecompendium te bouwen. Hierbij werd geopteerd voor een cross-platform compendium waarbij gegevens gemeten op verschillende microarray platformen worden gecombineerd in een uniform compendium. Dergelijk cross-platform compendium heeft als voordeel dat het een veel groter conditie bereik heeft dan de beschikbare 'single platform' compendia. Omwille van de heterogeniteit van de data gemeten op verschillende platformen werd veel aandacht besteed aan het ontwikkelen van de gepaste normalisatie procedures.

Met deze nieuwe methodologie construeerden we drie uitgebreide bacteriële expressie compendia nl voor *Escherichia coli*, *Bacillus subtilis*, en *Salmonella enterica* serovar Typhimurium. Bovendien werd een web portal met intuïtieve functies voor data-analyse en visualisatie ontwikkeld om een brede toegankelijkheid naar potentiële users te garanderen.

Een van de belangrijkste toepassingen van compendia is de studie van de conditie-afhankelijke respons van een organisme door het identificeren van conditie-afhankelijke coexpressie modules en door de onderliggende regulerende mechanismen te bestuderen die aan de basis liggen van de waargenomen expressievariatie. Hiertoe bestaan verschillende methoden, elk met hun

eigen assumpties en randvoorwaarden. In dit werk toonden we via een case studie (*sodA* gen expressie modules) aan hoe twee methoden DISTILLER en COLOMBOS complementaire resultaten opleveren bij het infereren van expressie modules. Door middel van dit voorbeeld, benadrukken we het belang van het kiezen van de juiste methode op basis van het onderzoeksdoel.

In een laatste fase hebben we de methodologie om compendia te genereren uitgebreid om ook een compendium op te stellen van een meer complex organisme i.e. *Zea mays*. Hierbij hebben we ons specifiek gericht op twee zaken, het oplossen van inconsistenties inplatform-probe relaties en het verkrijgen van een consistente conditie-annotatie die rekening houdt met verschillen in 'breeding lines', ontwikkelingsfases en weefselstructuren van maïs. Ook de web access portal werd uitgebreid met nieuwe functies aangepast aan queries die specifiek zijn voor een hoger organisme zoals *Zea mays*.

Abbreviations

ADF	Array Design Format file
BioCyc	Pathway/Genome Databases and Pathway Tools Software
BLAST	Basic Local Alignment Search Tool
CDF	Chip Description File
cDNA	complementary deoxyribonucleic acid
ChIP	Chromatin Immunoprecipitation
COLOMBOS	Collection of Microarrays for Bacterial Organisms
CORNET	CORrelation NETworks
DISTILLER	Data Integration System to Identify Links in Expression Regulation
DNA	deoxyribonucleic acid
EBI	European Bioinformatics Institute
EcoCyc	Encyclopedia of *Escherichia coli* K-12 Genes and Metabolism
FGS	Filtered Gene Set
GEO	Gene Expression Omnibus
GPR	GenePix Results format
GPR	GenePix Results

IDF	Investigation Description Format file
iPOP	integrative personal omics profile
MAGE-TAB	MicroArray Gene Expression Tabular
MaizeGDB	Maize Genetics and Genomics Database
MGED	Microarray Gene Expression Databases
MIAME	Minimum Information About A Microarray Experiment
NCBI	National Center for Biotechnology Information
NGS	next generation sequencing
PLEXdb	Plant Expression Database
PSSM	Position Specific Scoring Matrices
RNA	ribonucleic acid
SDRF	Sample and Data Relationship Format file
SOFT	Simple Omnibus Format in Text
UTR	untranslated region
ViTraM	VIsualization of TRAnscriptional Modules

Terminologies

Closed itemset	Frequent itemset that cannot be extended with an additional item without reducing the number of its occurrences
Coexpression	Genes' coherent expression responses to certain alterations in the organism's intra- or extracellular environment
Coexpression Module	A combined set of genes and condition contrasts, where the coexpression pattern appears
Compendium	Essentially an expression (log-ratio) value matrix, whose rows correspond to the known genes and columns correspond to sample contrasts
Coregulation	A phenomenon that genes are regulated by common transcription factor(s)
Frequent itemset mining	A data mining technology tries to extract information about a set of items that occurs together frequently
Homogenization	Our methodology to convert raw intensity expression values from individual biological sample into expression log-ratios between two different biological samples paired as a contrast so that they are comparable across different microarray platforms and experiments
Sample contrast	Where the expression values of genes from two biological samples, one set at the test and the other reference, are compared and the expression variations between samples are represented as log ratio

Sub-compendium	A subset of compendium containing contrasts sharing certain properties of interests, e.g., contrasts in which the gene expression profiles of different tissues are compared

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 High-throughput data and systems biology

Molecular biology has dominated biology research in the past century with exciting progresses: the discovery of DNA and its structure, the revelation of the central dogma of molecular biology, the disclosure of messenger RNA. Accompanied by the proliferate developments in biotechnology, such as recombinant DNA technology, polymerase chain reaction, and DNA sequencing, the processes of replication, transcription, and translation have been extensively studied. Due to the technological limitations, the molecular biologist has taken a reductionist approach to study a handful of genes or even a single gene in isolation to uncover its connections with the observable characters of an organism (phenotype). However, ever since the discovery of the complex regulation mechanism of the lac operon [67], there has been a growing understanding that an organism is a complex system of which characteristics and behaviors are often driven by complicated interactions among different components instead of a single gene. Yet, the early efforts were hampered by the lack of experimental technologies to harvest the data required to study an organism at the system level.

Heralding the coming high-throughput era, two technological breakthroughs in the 90s dramatically improved our data collecting ability, and revolutionized molecular biology research. First, automated DNA sequencers emerged and soon reached genome-scale sequencing [47], enabling, for the first time, the study of the complete genetic material and the discovery of the full gene universe of a species. This was followed by the development of the microarray technologies

[103, 111] that was utilized first to study the gene expression at genome-scale [77], and later to generate additional 'omics' data types [13, 129]. Due to their unprecedented ability, accompanied by the swift developments of required computational tools, these technologies were broadly adopted in biology research, and shortly in a scaled-up setting [120, 121]. The availability of various types of data for an organism at genome-scale finally expedited the emergence of systems biology as an approach for biological research. System level models of interacting components are constructed in order to provide insights into the function and control of biological systems that are not apparent from studying the individual component, and in turn, to generate experimentally verifiable hypotheses that deepen our understanding of those systems [61, 98]. The advance in systems biology commands ever more comprehensive data sets. However the amount of data generated by an individual experiment is physically constrained by the number of biological samples tested, whereas a massive volume of gene expression experiments are publicly available but largely remain underutilized. The research presented in this dissertation aims on bridging this gap by developing a methodology to construct comprehensive organism-specific expression compendia through a direct integration of publicly available experiment data.

1.2 Microarrays and data preprocessing

Fundamentally hybridization based, the microarray technology has its root in Southern blotting, where fragmented DNA attached to a substrate are probed with a known DNA sequence. Although the earliest approach that resembles microarray appeared in 80s [2], it is not until large amount of sequences became available following the advances in the automated sequencing technology that a microarray capable of probing the expression of thousands of genes simultaneously has become possible. And the first genome-scale microarray for *Saccharomyces cerevisiae* [77] appeared shortly after its complete genome sequence was released [54]. With its unprecedented ability, microarray has quickly become an indispensable technology to quantify global gene expression.

There are different microarray platforms for measuring gene expression, such as Affymetrix, Agilent, Nimblegen, or in-house microarrays [110]. They can be categorized according to the number of samples that can be hybridized simultaneous on a chip. *Two-channel microarrays* or *dual-channel microarrays* are typically hybridized with cDNA prepared from two samples and labeled with two different fluorophores (commonly Cy3 with green color and Cy5 with red color). Relative intensities of each fluorophore (color channel) are used to calculate ratios representing gene expression changes to identify up-regulated

and down-regulated genes between samples. A typical example of dual-channel microarrays is the cDNA array whose probes are cDNAs synthesized from mRNAs. In *one-channel microarrays* or *single-channel microarrays*, only one sample is hybridized to a chip generating intensity data for each probe or probeset. However these intensities do not reflect the absolute abundance levels of a gene but rather the relative ones when compared to other samples or conditions processed in the same experiment, as the experimental protocol and batch-specific biases render a direct comparison of gene's intensities of same platform origin across experiments uninformative. Affymetrix "Gene Chip", the most popular one of such platforms due to its high accuracy and precision [65], is used as an example to discuss the issues related to this type of arrays.

Microarray data are very noisy. There are many sources of systematic variation in microarray experiments that affect the measured gene expression levels, such as, heat and light sensitivity, dye labeling and detection efficiency, unequal quantities of starting RNA, etc [72, 140]. It is crucial to apply normalization procedures on raw expression intensities to remove systematic bias arising from the variations in the microarray technology rather than from biological differences between samples [105]. Various methods have been developed for this purpose, and most are platform dependent.

For cDNA microarray, the log-ratios calculated from the per-channel intensities from the same spot (probe) theoretically negate most of the systematic noises related to manufacture variations, such as, spot effect, print-tip effect, etc. However, often there exist intensity dependent artifacts in the obtained log-ratios which can be visualized in a MA plot[1] [138]. Locally weighted linear regression (LOWESS) analysis [24] first proposed by Yang *et al.* (2002) [138] has been shown to successfully remove such artifacts, and became the most widely used normalization method for cDNA microarray. This method was later replaced by LOESS (LOcal regrESSion) which is more flexible albeit computationally more intensive.

For single-channel microarray, such as the Affymetrix platform, a desirable normalization method needs to remove all variations of non-biological origin across arrays, as each array generates relative gene expression intensities for one sample only, whereas gene expressions of different samples (produced on different arrays) must be made comparable to observe true variations originating from biological differences between samples. Among the large number of methods developed for the popular Affymetrix platform, a global quantile normalization [14, 64] coupled with median polish summarization [62], combinedly known

[1] A MA plot is a visual representation of two channel DNA microarray gene expression data, which has been transformed onto the M (log ratios) and A (mean average) scale. In the plot, M and A values are represented in the vertical and horizontal axes respectively, and each point corresponds to one probe of the microarray.

as robust multi-array average (RMA) method, has been shown to outperform other methods [63] providing better precision even with few biological replicates. It soon became the *de facto* standard normalization method for this platform.

1.3 Gene expression compendium

As microarrays have become an indispensable technology for studying gene expression, a large amount of data has been generated. To promote data sharing, scientific journals generally require the deposit of the data results of these high-throughput experiments in public databases, such as Gene Expression Omnibus (GEO) [9] or ArrayExpress [101], upon publication. These databases are an extremely rich source of information, containing freely accessible data for thousands of experiments and a multitude of different organisms, and in theory provide an opportunity to analyze gene expression of a particular species at a global level [61]. Additionally, they hold the potential to expand the scope of any smaller scale study: mining the existing information offers molecular biologists the possibility to view their own dedicated experiments and analysis in light of what is already available.

1.3.1 Data integration issues

However, the opportunity of combining all public experiments for a single organism has not been explored due to two practical issues: data heterogeneity and data representation heterogeneity. First, microarray data are inherently highly heterogeneous. Data sets originate from different experimenters or labs and microarrays do not constitute a uniform technology. Even for data generated on one platform, protocols for sample preparation, labeling, hybridization, and scanning can vary greatly across labs and even studies, deteriorating the data compatibility [8, 65]. The consistency among data generated on different platforms is even worse due to the probe design and manufacture variations [35, 110]. Moreover, the aforementioned platform-dependency of preprocessing methods further lowers the data coherence, as often, different studies employ different methods [65, 115, 118]. Second, although community standards specifying the mandatory minimal experimental information accompanying each dataset (e.g., MIAME [16]) have been long established, the lack of the requirements [15] imposed regarding the format of the platform descriptions and the expression measurements, as well as the degree of preprocessing done on these values further complicates the matter of experiment integration from a practical point of view.

1.3.2 Existing compendium construction efforts

Despite such difficulties, several initiatives exist to actively build expression compendia from public resources. Reviewed in Fierro *et al.*, 2008 [46], the existing compendia took two different approaches to alleviate the aforementioned data integration issues: directly integrating data across experiments albeit limited by only those generated on a single-platform [42, 58], or combining results of individual experiment through meta-analysis to avoid direct data integration across experiments [29, 36, 70, 99, 106].

Meta-analysis is capable of combining data from different experiments across variant platforms, albeit in an indirect manner. It is generally a two step analysis: one first applies the desired analysis procedure (e.g., identifying differentially expressed genes, clustering gene expression profiles, etc.) on each single data set within the compendium separately, and subsequently combines the derived results. These compendia are often topic-specific, collecting all publicly available experimental information related to a subject matter of interest. ITTACA [36] and ONCOMINE [106], for instance, focus on cancer in human; Gene Aging Nexus [99] on aging in several species; GeneSigDB for gene signatures of cancer and related diseases [29]. Exceptionally, the ATLAS [70] initiative from ArrayExpress provides gene expression meta-analysis data sets for several species. However, containing only 400 experiments over all species in total and biased towards several eukaryotic model organisms, it is still far from comprehensive.

A direct integration approach removes data heterogeneity by applying normalization method across studies, then merges the results together to form one expression data set. Ideally, all data publicly available for a species could be integrated into one comprehensive compendium. Due to the aforementioned issues, the existing ones, however, focus on gathering only the data generated by a single platform, for instance, M^{3D} [42] or the commercial Genevestigator [58]. The Affymetrix platform is often chosen as the target platform, as they are the most widely used ones due to their robustness and high reproducibility [8, 65]. The single-channel nature of the Affymetrix platform and the use of proprietary file formats to report platform information and expression measurements avoid the data representation heterogeneity that plagues the data generated by many other platforms, and make the data collection task straightforward. Combining data from a single platform makes the in-between experiment normalization and probe mapping relatively straightforward, so that the quantitative measures of gene expression can be analyzed directly across experiments. For eukaryotic model organisms, such a single-platform approach works well as the compendium based on Affymetrix chips can achieve a broad scope on experimental conditions due to the platform's popularity. For instance, Lukk *et al.* (2010) [84] generated

a comprehensive data set for human containing 5372 microarrays representing 369 different cell and tissue types, disease states and cell lines. For prokaryotes, however, the single-platform constraint severely circumscribes the scope of the compendium created due to the lack of such a dominant platform. Even for a widely studied model organism, such as *Escherichia coli*, the most popular platform 'Affymetrix GeneChip E. coli Genome 2.0 Array' is used in less then one thirds of all publicly available experiments. When considering only one platform, a significant portion of data is missed out on. A novel approach that can integrate data of different platform origin is highly desirable.

When combining data from multiple experiments that address a set of related research questions, both meta-analysis and direct data integration approaches have the benefit of higher statistical power and more robust inference due to the increased number of samples in the combined data set and the reduced effect of study-specific biases [124, 132, 136]. However, when compared to direct data integration, meta-analysis approach has several disadvantages. First, the meta-analysis approach can only generalize those biological findings which are statistically significant in a high enough number of individual studies. This results in a great loss of information, and consequently leads to high false-negative rates [79, 137]. On the contrary, by analyzing a combined data set generated through a direct integration approach, new discoveries can be made that extend beyond the scope of individual experiment. Indeed, Warnat *et al.* (2005) [132] showed that direct integration reveals novel genes that are otherwise missed in single-set analysis, and the classifier incorporating those genes shows high predictive power and improved generalization performance. Similar observations were made by Fierro *et al.* (2008) [46], who compared meta-analysis with direct analysis of integrated data for differentially expressed gene retrieval. They noticed that when taking the results obtained by the analysis of a single experiment as a reference, direct analysis of multiple experiments detects more differentially expressed genes. On the other hand, indirect analysis tends to result in a rather restricted gene list which corresponds *grosso modo* to the intersection of the sets of differentially expressed genes obtained by individual experiment. Second, the compendia based on meta-analysis are limited by the predefined functionalities provided by the system, hence lack of flexibility to incorporate new analyses. The ones generated by direct integration, however, retain actual expression values, hence the compendia can be readily analyzed by the existing and future methods. This greatly broadens the scope of the utility of the compendium. For example, the *E. coli* compendium of M^{3D} [42], which was originally made to study transcriptional regulation in this species, has been applied to benchmark network inference algorithm [87], to discover conserved biclusters across multiple species [69], and to study chromosome conformation alternations under different physiological conditions [86], etc.

1.4 Objective and overview of the dissertation

Apparently, there exists a dilemma: on one hand the amount of data generated by individual experiments is physically constrained by the number of biological samples available, whereas on the other hand a massive amount of diverse gene expression experiments is publicly available but they largely remain underutilized. The main goal of this thesis aims to tackle this issue by creating a method that builds an organism-specific cross-platform gene expression compendium through direct data integration, and developing systems and tools that facilitate the access and utility of such compendium. Having the advantage of direct integration, while not being limited to a single platform, such compendium can provide an unprecedentedly broad coverage on varying experimental conditions and over diverse biological samples. To achieve the first goal, we carefully analyzed the issues that prevent direct data integration across platforms, then derived various specific strategies to handle those issues and conceived a three-step methodology for cross-platform compendium creation, and at last, developed a system incorporating the methodology to facilitate compendium creation and, vitally, the continuous curation to keep it up-to-date. The second objective focuses on identifying straightforward data analysis functionalities compatible with the type of data in the compendium, and developing a user-friendly system to facilitate the utility of data for a broader range of audiences. At last, we aim to expand the applicability of our methodology to eukaryotes by upgrading the existing systems with extra functionalities to handle the complexity of such species. This is demonstrated by creating a comprehensive *Zea mays* compendium.

The topics covered in this thesis are centered around this organism-specific cross-platform expression compendium (Figure 1.1). Chapter 2 describes the development of a methodology that enables the construction of an organism-specific cross-platform expression compendium. We first thoroughly investigated the issues involved in creating cross-platform compendium, including, the data representation heterogeneity, particularly for those related to the data generated on dual-channel microarrays, the lack of standards to specify experimental meta-data, and the sources of data inconsistency across experiments and platforms. Based on our study, we conceived a compendium creation methodology containing three major steps: data collection, annotation, and homogenization, each of which targets one specific issue identified. To reduce the complexity involved in compendium generation and to facilitate the maintenance of existing ones, we then developed a web system that provides user friendly interfaces to guide users through various steps of the compendium creation. The methodology lays the foundation of this research work. In Chapter 3, three such cross-platform compendia for bacterial model organisms (*Escherichia coli*, *Bacillus subtilis*,

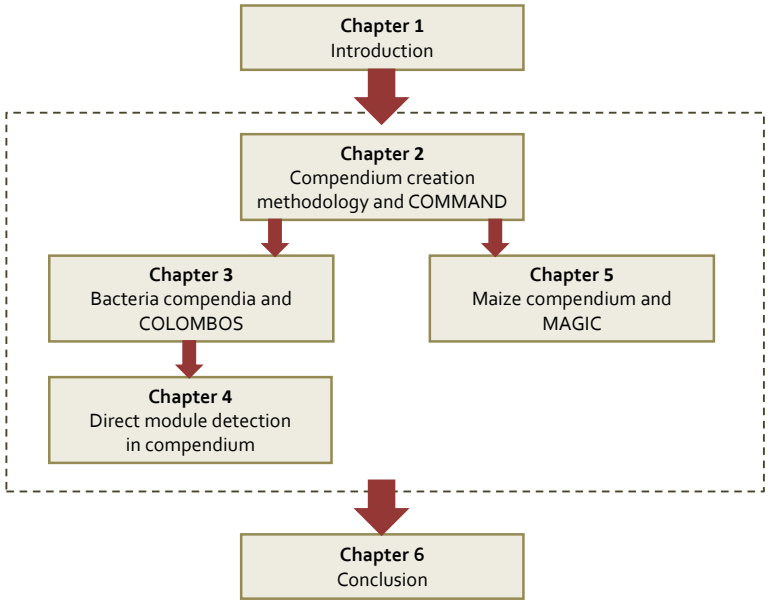


Figure 1.1: ***Organization of the dissertation***

.

and *Salmonella enterica* serovar Typhimurium) are presented together with a web access portal COLOMBOS that incorporates a suite of intuitive tools for data exploration, analysis and visualization and provides easy-of-use access to these compendia. The utility of both the compendia and the web portal is demonstrated in a case study, in which COLOMBOS analysis tools are employed to identify novel targets for *E. coli* transcription factor Fur (Ferric Uptake Regulation) based on their expression similarity to that of the known Fur targets across a range of diverse conditions in the *E. coli* compendium. Chapter 4 further showcased the utility of the compendium in the application of query driven condition dependent co-expression module discovery. Two web services, DISTILLER [80] and COLOMBOS [38], are recruited to explore the *E. coli* compendium and identify such modules for query gene *sodA*. We demonstrate that the complementarity between the results obtained by each approach well reflects the complementary nature of these two approaches, and the choice of the method depends on the nature of the biological questions to be answered. Chapter 5 describes our attempt to generate a cross-platform expression compendium for eukaryotic organisms utilizing the methodology. Here, we specifically address two issues: platform probe heterogeneity due to the

lack of the complete genome sequence, and precise biological sample annotation that manifests the abundant genetic repository (breeding lines) of maize species and its complex life style (development stage) and structure (tissue). At the end, the main achievements of this PhD study are summarized in Chapter 6 followed by some perspectives for future research.

Chapter 2

The cross-platform compendium creation methodology and COMMAND system

2.1 Introduction

Microarrays are one of the main technologies for large-scale transcriptional gene expression profiling. To promote data sharing, scientific journals generally require the deposit of these high-throughput experiments in public databases, such as Gene Expression Omnibus (GEO) [9] or ArrayExpress [101], upon publication. These databases are an extremely rich source of information, containing freely accessible data from thousands of experiments and for a multitude of different organisms, and in theory provide an opportunity to analyse gene expression of a particular species at a global level. Furthermore, they hold the potential to expand the scope of any smaller scale study: mining the information contained in such databases offers molecular biologists the possibility to view their own dedicated experiments and analysis in light of what is already available. So far, however, this wealth of public information remains largely untapped because these databases do not allow for a direct and integrated exploration of their data. The opportunity of combining the data from all public experiments for a single organism has not been fully explored due

to practical issues that can ultimately be attributed to the large heterogeneity inherent to microarray data. Data sets originate from different experimenters or labs and microarrays do not constitute a uniform technology. Multiple microarray platforms exist and are manufactured in different ways. Even for similar platforms, protocols for sample preparation, labeling, hybridization and scanning can vary greatly [72, 140]. Although community standards specifying the mandatory minimal experimental information accompanying each dataset (e.g., MIAME [16]) have been long established, the lack of the requirements [15] imposed regarding the format of the platform descriptions and the expression measurements, as well as the degree of preprocessing done on these values further complicates the matter of data integration from a practical point of view.

Despite such difficulties, several initiatives exist to actively build expression compendia from public resources. Most existing compendia can roughly be divided in two groups [46]: those that directly integrate single-platform experiments, and those that indirectly integrate cross-platform experiments. Combining data from a single platform makes the in-between experiment normalization and probe mapping relatively straightforward, so that the quantitative measures of gene expression can be analysed directly across experiments. Most single-platform compendia databases, such as for instance M^{3D} [42], or the commercial Genevestigator [58], focus on Affymetrix, one of the more robust and reproducible platforms [8, 65]. Combining data from different platforms, even to the extent of combining data from single- and dual-channel microarrays, is generally done by indirect meta-analysis as opposed to directly integrating the actual expression values: one first applies the desired analysis procedure (e.g., identifying differentially expressed genes, clustering gene expression profiles, etc.) on each single data set within the compendium separately, and subsequently combines the derived results. These compendia are often topic-specific, collecting all publicly available experimental information related to a subject matter of interest. ITTACA [36] and ONCOMINE [106], for instance, focus on cancer in human; Gene Aging Nexus [99] on aging in several species. There are exceptions though, such as the large ATLAS [70] initiative from ArrayExpress.

The compendia generated by directly integrating data across experiments have the advantage of retaining actual expression values, which broadens the scope of potential analysis procedures compared to indirect meta-analysis. Most of such compendia center on eukaryotic organisms, for which considerable amounts of data are available. Relying on only one platform can still lead to sizable compendia with a broad scope in condition content, such as the human compendium constructed based on the Affymetrix U133A platform with over 5000 samples [84]. However, for most species (e.g., *Zea mays*), no single

platform has such a dominant role. Even for well studied model organisms, such as *E. coli*, much less data are available on individual platform and a significant portion of the data is missed out when considering only one platform.

To have the advantage of direct integration, while not being limited to a single platform, and to facilitate compendium generation from public data, we have devised a methodology that directly combines expression data across platforms and experiments. The methodology has enabled us to create comprehensive compendia incorporating most high quality public data covering a broad range of experimental conditions as well as extensive types of biological samples across the boundary of experiment and platform. Although powerful the methodology is, to generate a compendium is still a complex and time-consuming task. To facilitate compendium generation and the continuous curation and expansion of the existing ones, a software system COMMAND (COMpendium MANagement Desktop) has been developed. The system guides user through every step of the compendium generation process with intuitive web interfaces. Although the complete options for each step are provided for advanced users, the merit of the system is that it enables quick sizable compendium creation through simple automated 'one-click' executions of the compendium generation steps, alleviating the complexity of this process. This has allowed us to generate multiple compendia that can be utilized not only to study a single organism but also to compare across species to study evolution and conservation.

2.2 Methods

2.2.1 Cross-platform expression compendium

The final goal of generating a cross-platform compendium is to generate a single data matrix that combines experimental results obtained from multiple microarray platforms performed under variant technical protocols. All genes measured by these microarrays should be represented, as well as all the experimental conditions that are under consideration. Differences due to technical features (platform, protocol, etc) should be removed to make the data comparable across experiments and platforms. The rows of the compendium correspond to the known genes of the organism in question constructed based on the corresponding RefSeq file at NCBI [104]. Uniquely, each column of it is a '*contrast*', which does not represent a single biological sample but the differences between a pair of samples, one as test and the other reference. Consequently, the expression values themselves are calculated as expression log-ratios representing the gene expression changes induced by the differences between this pair of samples. Converting absolute measurements of expression

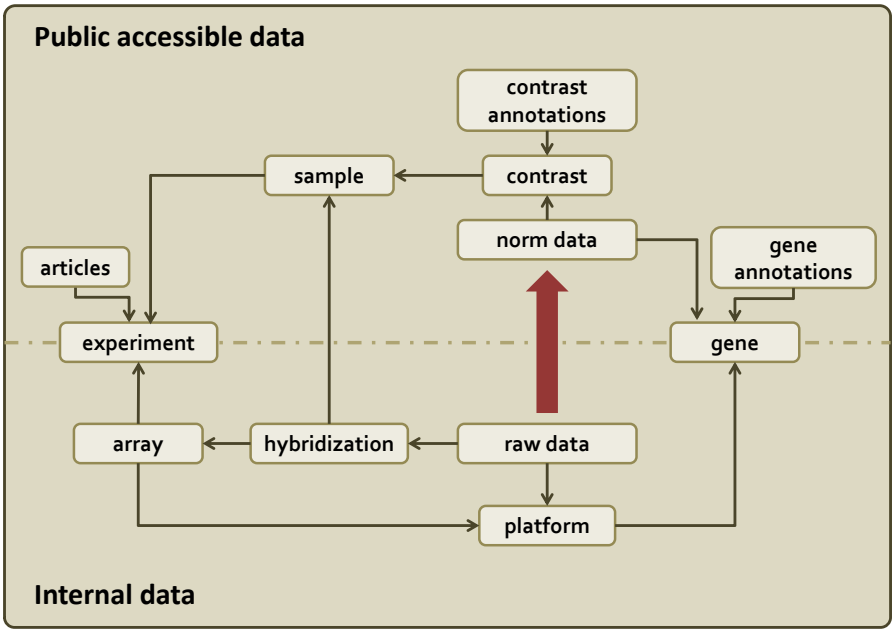


Figure 2.1: ***The database schema for expression compendium****. The schema is separated into two parts (by a dash-dot line). The entities in the lower part store the data that are collected from those online repositories. Containing various source of variations, these data cannot be used directly in biological studies. Instead, they are used to build the compendium, hence are internal. Those in the upper part store two types of data, the expression data that are generated from the internal data using our methodology and the annotations that are either manually curated by ourselves (for contrasts) or gathered from main curated external resources (for genes). These are the data that are explored and analyzed by the general users though the web portals desribed in Chapter 3 and 5, hence are the public accessible compendium data.*

into expression changes is the principal means for rendering expression values comparable across platforms and experiments. Relative expression calculated intra-experiment/platform (i.e. between two conditions measured in the same microarray experiment using one platform) negates much of the platform and experiment specific variations that make it impossible to reliably compare the absolute quantities reported in different experiments [115].

The database scheme for expression compendium

As the database representation directly reflects the different types of information and their relations that need to be extracted from public data to construct a compendium, we will briefly explain this below. The compendium is stored in a MySQL database. An abstract schematic representation of the various data and their relation is shown in Figure 2.1. The schema is separated into two parts. Those in the lower part store the original data that are collected from the public repositories and used only to build the compendium, hence are internal. Whereas those in the upper part store the compendium data that are publicly accessible through the web portals presented in Chapter 3 and 5.

Here we briefly explain each entity in the schema, which also relates to how the experimental data are organized conceptually. A gene expression *experiment* refers to a set of arrays that is designed to obtain expression data in order to answer a specific biological question. An *array* corresponds to one microarray in an experiment on which the biological sample(s) are hybridized to obtain gene expression measurements. A *hybridization* refers to an individual sample that is labeled and hybridized on an array. For an array performed on a single-channel platform like Affymetrix, there is only one hybridization per array, whereas there are two hybridizations, one per channel, for the array that were executed on a dual-channel platform. A *platform* denotes a specific microarray chip design with its corresponding probe annotation information, such as, the probe sequence, the physical layout, the target gene, etc. The *raw data* are the original measurements of each probe on a microarray chip. They are closed linked to the corresponding hybridization record in the compendium that reflects their sample origin. Three types of data are accepted: raw expression values, background intensities, and background corrected expression values. A *sample* refers to each individual biological sample that is used in an experiment. Biological replicates are treated as different samples. Note that a hybridization is an instance of a sample that is hybridized to a specific microarray chip. When the same biological sample is hybridized on several microarrays, they are referred as one sample but different hybridizations in the compendium. As explained in the previous section, a *contrast* specifies a comparison between a reference sample and a test sample, and the *norm data* (normailized data) are gene expression log-ratios. We use a relaxed definition for *gene*, which includes not only the protein coding sequences, but also other sequences that are expressed and functioning, given that their expression level can be quantified. There are also three supplementary entities: the article, the contrast annotation, and the gene annotation. The *article* stores the publications related to an experiment. The *contrast annotation* contains, for each contrast, the sample characteristics and the experimental factors that differ between the pair of samples belong to that contrast. The *gene annotation* stores the functional annotation of each gene

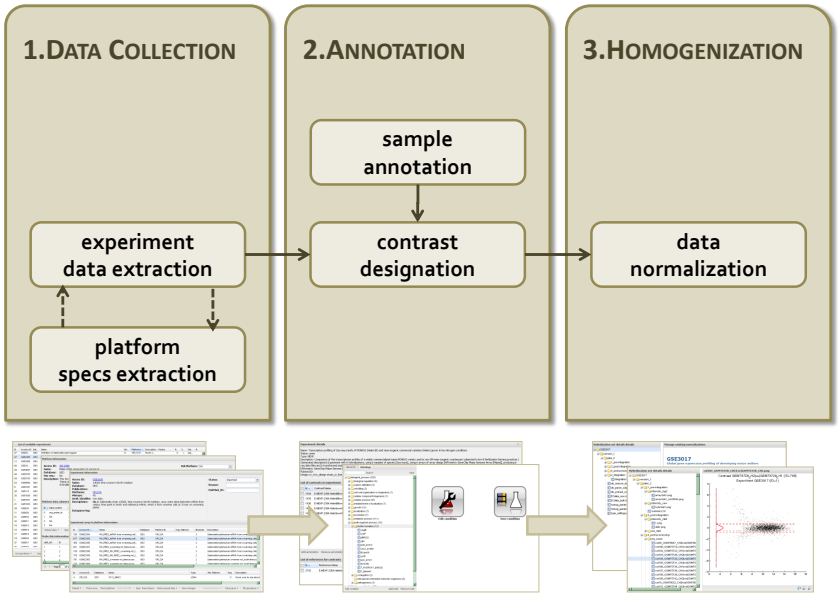


Figure 2.2: ***Cross-platform expression compendium creation methodology.*** *From left to right, the three boxes represent three main steps of the cross-platform expression compendium creation methodology. The corresponding functionalities are specified in the rectangles inside of each box. Below, the boxes shown the snapshots of the corresponding COMMAND interfaces.*

collected from external sources, including metabolic pathways, Gene Ontology annotation, transcriptional regulation, and transcription units. The contrast and gene annotations are stored to facilitate the querying and the biological interpretation of the expression data in compendium.

2.2.2 The compendium creation methodology

The cross-platform compendium creation methodology is composed of three major steps (Figure 2.2): data collection, experiment annotation, and data homogenization. First, in the data collection step, the raw expression data and the accompanying (experiment and platform) information are retrieved from the online repositories, overcoming the issue of the prevalent data representation discrepancies. Next in the annotation step, the contrasts are constructed

by assigning a pair of samples, one as reference and the other as test. The experimental factor and the characteristic differences between two samples are curated manually for each contrast and specified using a set of controlled in-house vocabularies. At last, in the homogenization step, the raw expression data are normalized and log-ratios are calculated based on the sample contrast designation to create the compendium data matrix. In the following sections, we will explain the individual steps in detail.

Step 1: Data collection

To generate a compendium, the gene expression data need to be collected from online public repositories, GEO and ArrayExpress. Although community guidelines such as MIAME [16] exist, the failure to provide a standard format to represent information of an expression experiment causes widespread representation discrepancies among the data deposited in those online repositories. The prevailing discrepancies coupled with the large amount of available data makes systematical expression data retrieval a daunting task. To make this task viable, we developed a semi-automatic workflow designed to tackle various data discrepancies and automate various processing procedures to handle the huge volume of data.

Here, we first explain the different sets of data to be retrieved, and then, discuss the data retrieval issues and the corresponding handling strategies for GEO and ArrayExpress separately, as each of them utilizes its own data reporting format and has its own specific data problems.

Data retrieval goal There are three sets of information to be extracted: the experiment metadata, the platform specification, and the raw expression values. The experiment metadata that describes the sample attributes and experimental factors driving the observed gene expression changes are essential for the data exploration and interpretation. Additionally, the metadata also specifies the relations between the samples and the microarray platform used, and the corresponding data tables or files containing the obtained values. As the expression values are measured by individual probes of a microarray, the reported measurements are related to probes rather than genes. The platform specification contains the probe-to-gene correspondence that is essential for converting probe measurements to gene expression values. The raw expression values are the foundation of the compendium. The quality of the compendium depends directly on the raw data's quality. The original probe intensities without background correction are the preferred type of the raw data, as the background correction has been shown to increase data variance especially at the low intensity range [63, 107], and the processed data contain artificially introduced inconsistencies due

to the different normalization algorithms applied. An in-house normalization pipeline is utilized on the collected raw data to achieve better consistency. Except for extracting desirable data, equally important is to retain proper value-to-probe mapping and value-to-hybridization association (see Figure 2.1). The former is required to convert the raw probe intensity into the gene expression value. The latter links the expression value with the corresponding biological sample, hence the associated sample attributes and experimental factors.

Step 1.1: Experiments retrieval, filtering, and data downloading Although started with gene expression data, GEO and ArrayExpress are extended through the years to include other functional genomics data, such as, ChIP-chip experiments. Furthermore, certain types of expression data, e.g., those of the cross-species gene expression comparison studies, are not interesting for a compendium specific for one species. It is then preferable to start with a clean list of experiments.

To maximally automate this process, the programmable access interfaces, Entrez Programming Utilities (E-Utils) of GEO and REST query API of ArrayExpress, are utilized to retrieve the basic metadata of all experiments belonging to a given species. Note that GEO, as the earliest of these two, serves as our primary data source, whereas ArrayExpress supplements GEO by providing extra data not available in GEO. Upon obtaining the list, the existing experiments are removed. For the experiments common in both repositories, only the corresponding GEO record is kept. The remaining experiments are sifted through filters based on a set of in-house collected key phrases to remove undesirable ones, e.g., ChIP-chip data, cross-species comparison data, etc. The filtering is rather conservative to avoid excluding experiments by mistake. The remaining experiments are then reported as new ones. A manual inspection on this list of new experiments is recommended, although not mandatory. When definite, the corresponding data file(s) are then downloaded from online repositories for data collection.

Step 1.2: GEO data extraction Here we first explain how the expression data are organized in GEO and then the strategy to extract data desirable for the creation of the compendium.

Expression data in GEO There are two sets of data for an experiment in GEO, the *GEO Series* (GSE) containing the original submitted data and the *GEO DataSet* (GDS) containing curated normalized data. To generate the compendium, we collect original data stored in GSE. The data of an experiment is stored in a standard simple line-based plain text format, called SOFT (Simple Omnibus Format in Text). There exists one SOFT file per GSE (experiment) stored in compressed (gzip) form. The file contains one or multiple instances of three types of records: Series, Sample, and Platform (see Table 2.1). The *Series*

Table 2.1: GEO Records en Files and their corresponding content

Record	Compendium entity	Information contained
Series	Experiment	experiment metadata
Sample	Array	sample characteristics, hybridization(S) (channel) information, and data table (expression measurements)
Platform	Platform	platform metadata and probe annotations
Supplementary files	Array or Hybridization	raw expression values accompanied with probe annotations in vendor specific format

record contains general information about various aspect of an experiment described in free text. The references to the corresponding Sample and Platform records are also provided. Each of the *Sample* records corresponds to one microarray chip used in an experiment. It is composed of two sections. The first one provides free text descriptions about biological samples hybridized to the microarray, including information about sample attributes and experimental factors crucial for compendium data exploration and interpretation. The referral to the related Platform record is also provided here. In the second section, the expression measurements are reported for each probe (or probeset) of the microarray in a tabular form. Often the values reported here are processed data instead of the desirable original probe measurements. The *Platform* record describes the microarray chip used to measure gene expression, including a free text description and its probe annotations in a tab-delimited table. Specific about the required content in various records and their sections, the SOFT format does not impose strict requirements about the representation of the data. Consequently, the expression data reporting format and the probe annotations vary between experiments and platforms. For many experiments, raw data files are also available as supplementary files providing an alternative source to obtain desirable gene expression measurements. In original vender specific formats, the raw data files differ among software and/or equipment. Often allowing to be used alone, the file includes probe specification for each record in addition to the expression value(s). Note that as the submission of the raw data files is optional, many experiments do not have supplementary files. Each record in GEO has a unique access id. Its 'Series', 'Sample', and 'Platform' records correspond to the experiment, array, and platform object in our compendium respectively (see Table 2.1 and Figure 2.1).

By default, to obtain expression data from a GEO experiment, the corresponding SOFT file is analyzed. The contents of different records contained in the SOFT

file are separated first, and then handled by the specific parsers to extract data. Only when the desirable data cannot be obtained, the corresponding supplementary files are processed.

Step 1.2.1: GEO experiment metadata extraction The experiment metadata includes the basic experiment information in a GEO 'Series' record (e.g., name, description, publication, etc) and the sample attributes and experimental factors in the corresponding 'Sample' records. As mentioned before, instead of using controlled vocabulary, this set of information is described in free text, which lacks consistency as often one concept can be described in multiple ways. Additionally, often the metadata available is incomplete, and the missing information needs to be manually extracted from the related articles. The inconsistency and incompleteness of the experiment metadata renders a computerized data extraction impossible. Instead, a manual curation (in the annotation step) is required to analyze and standardize the information into a computable format to facilitate data exploration. Consequently, this information is extracted as is without modification. And the direct correspondence between those GEO records and the compendium objects makes the extraction straightforward.

Step 1.2.2: GEO platform data extraction The GEO 'Platform' record contains both the basic platform description and the probe annotations. The former, including name, description, manufacturer, etc, are extracted as is. Although the probe annotation is always presented in the form of a table, each platform has its own format that varies in the amount of content (number of columns), the column naming convention, and the content of each individual column. Ideally, the probe sequences should be provided so that a homology search against the latest gene sequences using BLAST [1] can identify the up-to-date gene target for each probe. However, although required by the MIAME[16] standard, this information is missing for the majority of the platforms. Alternatively, the target gene can be identified by other information, namely, locus tags, alternative gene tags, or common gene names. But the lack of a consistent format makes it hard to identify the column providing the proper information. Even worse, sometimes, the relevant information is embedded in a column in which multiple contents are specified in a complicated structure. Additionally, the inconsistency also exists in the content of one data column, in which different types of information are found. For example, locus tag, gene name, or other tags are often used alternatively in one data column titled 'ORF'.

The key for platform data parsing is the ability to identify data columns containing useful information and provide proper methods to extract a standardized set of information for every platform. Due to the format heterogeneity of platform data, a user-guided semi-automatic procedure that couples manual data column identification with the automated data extraction

is developed. First, the probe annotation table is checked manually to identify data columns containing desirable information and specify the method to extract data. While collecting platform data for parsing, an in-house dictionary based content identification method is applied to analyze the column names of the annotation table and automatically mark out the candidate columns. Helping to alleviate the manual inspection task, it cannot, however, replace the human inspection, as it cannot handle information embedded in a complex format and the column name based identification is not 100% reliable. Even with data columns identified, the existence of different formats that can be used to specify one type of content complicates data parsing. To automate the content parsing task, a plug-in based system with great flexibility and extensibility was developed. The existing nine plug-ins together with the direct 'copy' option are capable of handling every platform included in the compendia already developed, and the system can be easily extended with new plug-ins. Next, after manually identifying interesting data columns and assigning proper function to parse them, the data are extracted utilizing our plug-in system, and the target genes are automatically identified. When identifying target genes without sequence information, the aforementioned content inconsistency is handled by an integrated search over multiple information sources in strict order of preference, locus tag, alternative gene tag, and then common gene names, based on its reliability.

Step 1.2.3: GEO raw value data extraction As mentioned previously, the expression value extraction has two goals, to extract desirable raw expression data and to retain proper value-to-probe mapping and value-to-hybridization association. In GEO, the expression data can be extracted from two different sources, the 'Sample' record and the supplementary file. The former is the primary data source, as it is readily available in SOFT file. The supplementary file is checked only when the desirable data cannot be extracted from 'Sample' records. The expression value extraction for the single-channel array is easier than that for the dual-channel array, as the value-to-hybridization association is straightforward, given that only one hybridization exists. To appreciate the full complexity of the expression value extraction task and to demonstrate the full capacity of our strategy, we assumed that the data to be handled are generated by a dual-channel microarray, whereas the single-channel microarray data can be handled with the same strategy by leaving out the value-to-hybridization association extraction part.

Step 1.2.3.1: Parse GEO Sample record Each GEO 'Sample' record has a data section that reports expression value in a tabular form. Although only processed results are required to be reported here, often they are accompanied by the corresponding raw intensities. As part of the SOFT format, the proper value-to-probe mapping is guaranteed because the same probe identifier are used

consistently across the corresponding 'Sample' and 'Platform' records. This simplifies the data extraction task. Hence it serves as our primary source for raw data extraction.

To successfully extract raw data, two tasks remain: to properly identify the desired raw expression values and to obtain correct value-to-hybridization associations. The first task is hindered by several issues. First is the lack of standard data reporting format in 'Sample' record, instead the data table of the corresponding raw data file is often used. The existence of a large number of raw data formats creates widespread data representation discrepancies. Although it is mandatory to provide the data table header descriptions explaining the content of each data column, depicted in free text, it cannot be readily analyzed by a computer. Secondly, measuring two samples in one chip makes it crucial to identify the channel association for each data to be extracted. However this information can only be derived indirectly through analyzing the column name. At last, sometimes, only the background corrected values instead of the original ones are reported. As the correction results in an increased variance for lower, less reliable intensity levels [107], these data are not desirable. When the corresponding background value is available, the uncorrected intensity can be reconstructed from the data. However, this requires that the system is capable of automatically identifying and applying certain data conversions. Note that since raw intensities are reported as numbers, the data extraction is straightforward after the corresponding columns are identified. For the task to obtain value-to-hybridization associations, the channel related information extracted from column name needs to be paired with the hybridization information specified in the 'Sample' record. Although 'ch1' and 'ch2' are used in the 'Sample' record, various types of information are specified in the data column names.

Given the above issues, we developed a semi-automatic raw expression value extraction system. The fully automatic extraction is triggered under a strict condition, in which the raw expression value without background correction and the corresponding background values are extracted, the channel information of each extracted data is properly identified, and the value-to-hybridization associations are correctly derived. Such a strict condition guarantees that correct data are extracted by the system. The key to the automatic data extraction lies in the ability to programmatically identify the data content type and the channel information, avoiding manual inspection. Our solution is a rule based column name analysis subsystem that analyzes the name of data column using pattern matching to identify required information. Targeting the desired data, the system focuses on identifying three types of information, the raw expression intensity value (I), the background value (BG), and the background corrected expression value (I_{BG}). For each data type, the related columns are collected from large number of Sample records. The patterns of the column

names are then manually analyzed, and the rules to recognize the data type and the channel related keywords are derived and used in the system. Next, the extracted channel keywords are checked against the hybridization specification to obtain value-to-hybridization association. Three types of keywords are handled, the direct channel specification ('ch1' and 'ch2'), the dye color specification ('cy3' for green and 'cy5' for red), and the dye frequency specification ('532' for green and '635' for red). The first type can be directly mapped to a hybridization, whereas for the last two types of keywords, the hybridization labeling information is required. When all columns are analyzed, the system checks those identified and automatically discovers and sets the data conversion function when necessary. At last, the aforementioned conditions are checked against the information identified by the system. When successful, the raw expression data are then automatically parsed. Otherwise, the corresponding error is reported to highlight the issues for manual inspection.

Step 1.2.3.2: Parse GEO supplementary file When the 'Sample' record does not provide desired raw values, it is still possible to extract them from the raw data file that supplements the 'Sample' record. Note that designed to be self-contained, the raw data file includes not only the expression value data, but also the probe annotations and various levels of meta information. Except for the same issues encountered when parsing 'Sample' record, there are other issues when parsing raw data files. First, as mentioned before, there exist a large number of different formats for raw data files, which varies on the mount of the meta information available, the probe annotations used, and how the raw values are reported. Hence, each format needs a specific parser to handle it, and to automate the process, the system should be able to automatically recognize the file format. Second, as the raw data file is not part of the SOFT format, the probe annotation specified in the file does not use the same GEO probe identifier specified in the corresponding 'Platform' record. Whereas to improve the consistency across the experiments using the same platform, the GEO probe identifier is preferred, and the probe information in a raw data file needs to be mapped to this identifier. It should be noted that the reporting format used by raw data file is not platform specific, hence the parser should be flexible to handle different platforms. At last, the relationship between a 'Sample' record and its corresponding raw data file is not always reported, hence the system should have the capability to identify it.

To guarantee the extraction of correct data with proper probe and hybridization associations and to automate the parsing process, several format specific parsers are developed to handle most popular raw data file formats. Currently, following formats are supported, GenePix Results format (GPR), and Perkin-Elmer ScanArray format. However, the NimbleGen Pair format and Imagene format is

not supported, as manual intervention is required to handle them[1]. Each data format is thoroughly studied to develop the corresponding parser. Similar to 'Sample' record data parsing, the raw value data columns and the corresponding value-to-hybridization associations are identified from column names. The focus of a parser is to provide a format specific method to obtain accurate value-to-probe mapping, which should be flexible to handle different platforms. A greater accuracy is achieved by utilizing the well structured and accurate parsed probe data of the corresponding platforms to search against the standardized probe information provided in a raw data file. A match could be identified in multiple manners applied in the order that favors the most accurate one applicable, providing flexibility to handle the amount of information that is different between platforms. Consequently, to parse a supplementary raw data file requires that the corresponding 'Platform' record has been parsed obtaining as much information as possible for probe matching.

The supplementary file is parsed by a semi-automatic three-step procedure. First, the raw data file to GEO 'Sample' record association is either obtained from the 'Sample' meta data or derived from the existence of the GSM access id (identifier) in the name of the corresponding file. Next, the format of the raw data file is identified by either the file extension or its meta information signature. When identifiable, the corresponding dedicated parser is then applied to automatically extract raw values and assign them to the proper probes and hybridizations. For the file of unknown format, a manual data parsing is provided as an alternative. The new formats encountered are tracked, and a dedicated parser can be added when necessary. To avoid overcomplexity, the automated system only handles the case where there is only one raw data file per 'Sample', and supports only the ASCII file and the excel binary file (csv or xls).

Step 1.3: ArrayExpress data collection

Expression data in ArrayExpress In ArrayExpress, the experiment data are stored in the MicroArray Gene Expression Tabular (MAGE-TAB) format. In this format, 4 types of file capturing different sets of information are used, namely, the Investigation Description Format (IDF) file, the Sample and Data Relationship Format (SDRF) file, the Array Design Format (ADF) file, and the raw and processed data files (normally packed into one zip file) (see Table 2.2). Similar to the 'Series' record in GEO SOFT format, the *IDF* file provides an overview for an experiment. Although the MGED ontology terms are used, they specify only the type of information. The content, however, is still described

[1] The results of a dual-channel platform reported in these formats are split into two data files, one per channel. It is often impossible to programmatically derive the value(file)-to-hybridization association.

Table 2.2: ArrayExpress data file format and corresponding content

File	Compendium entity	Information contained
Investigation Description Format (IDF) file	Experiment	experiment metadata
Sample and Data Relationship Format (SDRF) file	Hybridization	sample metadata; links between sample, array (indirect), and data file
raw data file	Raw data	expression values
Array Design Format (ADF) file	Platform	platform metadata and probe annotation

in free text. The *SDRF* file describes the sample characteristics and the relationship between samples, arrays, and data files. The content is presented in a tabular form, in which each record (row) describes information about one biological sample hybridized to a specific channel of a specific microarray chip, hence corresponds to a hybridization in our compendium (Figure 2.1). The references to the corresponding platform are provided for each record. One caveat about the SDRF file is that its format is not strictly defined, and can vary among experiments. Next, the *ADF* file provides information for a microarray platform, including both descriptive information and the probe annotations. At last, the expression values are reported in the raw and/or normalized data files. In ArrayExpress, the unique access id is given only to experiment and platform, which directly corresponds to the experiment and platform object in our compendium (Figure 2.1). However, there is no explicit record in ArrayExpress that corresponds to the array object. Additionally, although each record in SDRF file corresponds to the hybridization object, without a unique reference, this information is only indirectly accessible through the corresponding ArrayExpress experiment.

For an ArrayExpress experiment, the corresponding IDF and SDRF files together with the raw data file are downloaded. The first two contain experiment metadata, and the last one contains the expression values. Additionally, for each new platform, the corresponding ADF file is downloaded and handled separately.

Step 1.3.1: ArrayExpress experiment metadata extraction For an ArrayExpress experiment, the experiment information is described in the IDF file, whereas the sample attributes and experimental factors are described in the SDRF

file. The metadata extraction from the IDF and SDRF files is not so simple. The first issue is the data content inconsistency in SDRF file, as it has no strictly defined format. The data can be submitted into ArrayExpress using different methods, e.g. MIAMExpress, MAGE-TAB, etc, generating compatible but not identical sets of information. For example, 'Hybridization'[2], which is mandatory in MIAMExpress, is replaced by 'Assay' in MAGE-TAB, which is not obligatory and often missing. Secondly, there is no information that directly specifies the individual microarray chip (corresponds to an array record of the compendium) used in an experiment and the corresponding channel(s) of each chip. For a dual-channel microarray, it is crucial to pair two channels together, as their raw data are often jointly normalized. Recall that each SDRF record (row) corresponds to one channel of an array, this relation between channels then needs to be derived indirectly from the data content. Due to the data inconsistency, multiple data sources (columns), e.g. 'Array Data File', 'Scan Name' or 'Hybridization Name', are utilized in this process, depending on their availability. When this relation is identified, the corresponding SDRF records are paired and assigned as different hybridization(s) of a compendium array object. When the system fails to identify this relation, a manual inspection is required.

Step 1.3.2: ArrayExpress platform data extraction For each platform in ArrayExpress, the data are specified in the ADF file, which needs to be downloaded separately. Similar to the GEO 'Platform' record, there are two sections in an ADF file, the platform descriptions and the probe annotation table. However, the platform probe annotation table in ADF file follows a standardized format with a fixed set of column names and rather simple data content without complex structures. An automated data extraction method targeting the preselected data columns that provides desired probe data is used to parse the ADF file, avoiding the issue of the data column content type identification that plagues the GEO platform parsing. Although the content inconsistency still exists, it can be handled using the integrated search strategy developed for GEO platform data parsing.

Step 1.3.3: ArrayExpress raw data extraction In ArrayExpress, the expression values are reported in the raw data files and normalized data file. The raw data files are our focus as they provide the data for creating a compendium. There exists two kinds of raw data files, the vendor specific data files, e.g. Affymetrix CEL file, Genepix GPR file, etc, and the processed raw data files, which are normally reformatted. The former are used mainly for the experiments based on

[2]Noted that this 'Hybridization' is similar to the GEO 'Sample' record, which refers to one microarray chip used in an experiment and corresponds to the compendium array object. This is different from the compendium hybridization object that corresponds to one channel of an array.

Affymetrix platforms, and occasionally, for the experiments using GPR files. The Affymetrix specific data files (CEL file) are handled by a dedicated procedure that will be explained later. The experiments using original GPR files are very rare. Only two such experiments were found while creating all four existing compendia. As the probe information specified in GPR file could not be mapped to the corresponding ArrayExpress platform's ADF specification without loss of information, they are skipped for now. The majority of the experiments, including most of those originally using GPR files, report expression values using the processed raw data files, which is a tab delimited text file containing only an adapted data table without meta information. Two kinds of changes are applied on the original data table. First, the platform related part of the table, containing probe annotations, is replaced with the corresponding standard ArrayExpress ADF platform annotation. Consequently, it is straightforward to correctly obtain value-to-probe mappings between the raw data file and the ADF file. Second, the column names of the expression data section of the table are augmented with different types of information. The most common ones include but are not limited to the original file format, the corresponding 'Hybridization Name' specified in SDRF file, or the original data file name. In a majority of cases, this feature allows a separation of the platform related data from the expression value related data, and the recovery of the original column names.

The raw data extraction is proceeded only when it is possible to link the extracted expression values to the corresponding biological samples. Note that the SDRF file parsing has successfully reconstructed channels relations and created the corresponding compendium array object. However there still exist two missing links to connect raw data to sample. The first link is to identify the corresponding raw data file for each array object. Generally, the 'Array Data File' column in SDRF record specifies this information. When that is missing, we noticed that the extra information added into data columns sometimes provides clues, for example, when it matches the hybridization name of a SDRF record. In those cases, the system is capable of automatically identifying the data file correspondence. Otherwise, a manual inspection is required when this correspondence cannot be identified. The second link is to obtain the value-to-hybridization associations for the data generated on dual-channel microarray platform. As the original raw data column names can be recovered, the same rule based column name analysis subsystem developed for GEO data parsing is utilized to extract the dye color information from column names, which is then checked against the 'Label' of a hybridization obtained from SDRF record to recover the required associations.

Hence the ArrayExpress processed raw data file are parsed in five steps. First, the data file to compendium array object association is checked or identified.

Next, the data file headings are analyzed to separate the probe annotation columns from the raw data columns. Then the raw data column names are analyzed to recover value-to-hybridization association. At last, the same set of data quality conditions applied in GEO 'Sample' data parsing are checked. When met, an automated data retrieval is executed. It extracts, for each hybridization, the raw values from the corresponding channel and also maps each value to the correct probe. Similarly, when it fails, the corresponding error is reported to highlight the issues for manual inspection.

Step 1.4: Affymetrix platform data and expression value extraction The single-channel Affymetrix microarray is, by far, the most widely used platform due to the high consistency among the results obtained across labs compared to other platforms [65]. The use of the proprietary file formats to specify platform information (Chip Description File, CDF) and to report expression values (CEL file) simplifies the data retrieval.

In an Affymetrix chip, each gene is targeted by a group of short oligonucleotide probe pairs, collectively called a *probeset*. Each probe pair is composed of a Perfect Match (PM) probe and a MisMatch (MM) probe, in which the PM probe measures the target gene expression level and MM probe measures the background signal. Due to this specific design of Affymetrix microarray chip, there exist two possible platform specifications, one for probesets and the other for probes. The former is reported in GEO and ArrayExpress, whereas the latter can be retrieved from the corresponding CDF file. Consequently, two levels of expression value exist, the raw intensity of each probe, and the summarized intensity of each probeset. Many algorithms [59, 62, 83] have been developed to normalize Affymetrix data and compute the summarized expression value for each probeset. To avoid the inconsistency introduced artificially by different algorithms, we opt to obtain the raw probe intensities from the CEL file, then processed using our in-house homogenization pipeline based on RMA algorithm. The background values measured by MM probes are ignored as it has been shown that, though the background correction improves accuracy, it greatly sacrifices the precision[63].

Occasionally, an experiment using Affymetrix platform reports only the summarized expression values in the online repository. This requires that our system is able to handle both raw intensities and the summarized values, and in turn, requires retaining both the probeset and probe annotations, and the relations between them. GEO and ArrayExpress only store the probeset annotations of an Affymetrix platform. To keep it consistent, the corresponding compendium platform record contains the same annotations. Additionally, for each Affymetrix platform, an extra platform record, which is always associated with the original one, is created artificially in our compendium. This kind of

platforms are called 'virtual platform' as there exists no directly correspondence in the online repositories. Next, by incorporating the proprietary Affymetrix Fusion SDK into our system, the probe specifications are extracted from the corresponding CDF file downloaded from Affymetrix website and stored with the virtual platform.

With platform specification ready, the raw intensities can be easily extracted from the CEL file. As a single-channel platform, each CEL file contains the expression values for only one hybridization (and one sample). For an experiment for which the CEL files are available and the file-to-array correspondence has been successfully derived (using aforementioned repository specific methods), the raw intensities can then be automatically extracted from CEL file using Fusion SDK. As both CEL and CDF files are Affymetrix proprietary formats, the same probe references are used consistently. At last, the corresponding compendium array object is linked to the virtual platform instead of the original one to reflect the fact that the probe level raw intensities are retrieved as raw data instead of the summarized values. If the CEL file is missing, the summarized values are used.

Step 2: Annotation

After obtaining the experiment data from online repositories, the contrasts that will be represented in the compendium are defined. The experimental factors and the sample attributes are carefully studied to construct a rigid annotation for each contrast.

Contrast designation Based on their biological role in an experimental survey, hybridizations are labeled as 'reference' or 'test' on a per experiment-and-platform combination basis and matched to produce a set of contrasts.

For dual channel experiments, usually one of two hybridizations of an array serves as a reference to the other, as this inherently counters much of the probe associated variation in the measurements. There are exceptions however, such as when one of the hybridizations on an array does not constitute an identifiable and unique biological condition for which the transcriptome was assessed (e.g. a sample of genomic DNA or a pool of different samples that cannot be considered as biological replicates). These hybridizations are discarded and the experiment is further treated as if it was a single channel experiment. In this way we ensure that every contrast has a biologically interpretable meaning: its associated log-ratios represent expression changes in response to quantifiable stimuli altered from reference to test or the characteristic differences between samples.

For a single channel experiment, one or more hybridizations can be chosen as references for the remaining tests or each group of tests respectively. The choice depends on the nature of the experiment and the type of biological condition that is measured. Just to give a few examples. For time-series experiments, the sample taken at the first time point is chosen as the reference to those taken at the other time points. For an experiment studying the effect of a set of mutants against the wild type, the latter is chosen as the reference. A more complex example where we would choose more than one reference might be an experiment where the transcriptomic responses to the treatments with two compounds A and B were measured at different concentrations. In such a case, we use the treatment with the lowest concentration of A as the reference for all other treatments with A, and the treatment with the lowest concentration of B as the reference for all other treatments with B, and we would additionally include one contrast comparing both references, i.e. where the lowest concentration of A is considered test and the lowest concentration of B is considered reference (to include an explicit comparison between A and B). It is important to note that for an experiment using several platforms, the samples are grouped by the platform and for each group, one sample is designated as the reference for the other samples in the same group. This is because the probe differences between platforms render the measured gene expression intensity values incompatible. Consequently, it unjustifies the use of a sample measured on one platform as the reference for the samples measured on the other platforms.

Multi-chip platforms Microarray can often not cover the complete gene set of an eukaryote in one chip. Multiple chips of the same technology, each targeting complementary gene sets, are therefore needed (referred as the *multiple-chip platform*). Ideally, the data generated on multiple chips originated from the same sample should be grouped together by assigning corresponding hybridizations to the target sample in the compendium. To this end, a special handling strategy has been developed to properly normalized data generated on this type of platform (section 2.2.2).

Sample annotation and ontology Given a set of formal hierarchically structured properties, we can use the differences of these property values between the test and reference samples as the annotation. For the contrast annotation in bacteria compendia, four classes of properties are defined: genomics, growth, medium, and treatment. The properties of the class 'genomics' specify the genomic difference(s) of respectively the test and reference samples. It contains four subcategories: mapped mutations, phenotypic strains (carrying specific phenotypes with yet mapped mutations), evolutionary adaptation (genomic differences accumulated in an evolutionary experiment), and plasmid (genomic differences in plasmid DNA). The differences in the chemical compounds and additives used as the media of respectively the test and the reference are

described as the medium related properties. The class 'treatment' includes differences bewteen test and reference in general environmental properties, such as, temperature, pH, UV radiation, or oxygen level etc. Differences in other general properties are grouped as growth properties, such as, time, growth phase, etc. Each annotation specification is defined as a duplet, including a property it is different between the samples of a contrast and a value describing the extent of this difference. The annotation consists of a vector of values, one for each characterized difference between samples. Our annotation enables a mathematical comparison and automatic organization of contrasts based on the properties that are surveyed, but it is a labor intensive manual curation process where information often needs to be retrieved from original publications, supplementary data and occasionally directly from the authors.

The annotation properties themselves are further structured in an ontology tree employing the same classes as the well-defined Gene Ontology biological process subtree terms [53]. Assigning annotation properties of seemingly distinct categories to the same ontology term reflects the fact that different properties might in fact be related to the same biological process. For example, in the *E. coli* compendium, the condition ontology term 'response to oxygen levels' includes several properties from different levels of the property hierarchy, but that are all linked to cellular processes dependent on oxygen availability, such as *fnr* mutations (a global oxygen responsive transcriptional regulator), NO2 concentration (an electron transport decoupler), agitation of the growth medium, actual oxygen levels, etc. The ontology provides a biologically intuitive view of the annotation, and a novel data exploration option allowing an integrated study of different aspects centered around targeted biological processes.

Sample annotation for eukaryote To handle the biological diversity of the complex plant *Zea mays*, the annotation system is further expanded to completely characterize, for each sample, its genotype specification (breeding line), tissue origin, and the development stage. Associated with each sample, these characters extend the scope of the annotation to provide extra information that might be common between the test and reference samples, yet crucial for data exploration and interpretation. For example, the expression variation observed under a stress condition when comparing a pair of leaf samples might differ from that between two root samples.

The detailed sample characterization through the extended annotations enables a refined categorization of contrasts based on those characters. Four sub-compendia are created containing only a subset of contrasts that compare between samples gathered in different tissues, collected at variant development stages, taken from distinct breeding lines, or with or without external perturbations respectively. Providing a sub-compendium with a confined scope

could well facilitate the data exploration for the biologist with particular research questions. More details are given in section 5.2.2 of chapter 5.

Step 3: Data homogenization

The final part in the creation of a compendium is the homogenization in which raw expression data are normalized, and the log-ratios are calculated, per experiment, between samples based on the contrast specification. Several preprocessing procedures are conducted to render expression levels comparable between different experiments and platforms. Crucial steps in this preprocessing are array-specific and depend on both the technological platform used to perform the experiment (single- or dual-channel), as well as on the reported units of expression. In general we adhere to the following principles:

1. Raw intensities are preferred as data source over normalized data provided by the public repository. (This is handled in the data collection step.)

2. No local background or mismatch probe correction procedures are performed to avoid an increase in variance for lower, less reliable intensity levels [37, 63, 83, 107]. This improves the data precision, which is crucial for our compendium, since we do not do any 'significantly differential expression' calculations, which might be robust against the increased variance, due to the lack of necessary biological replicates.

3. Non-linear normalization techniques are performed to account for global inter-hybridization differences (e.g., a loess fit to remove dye-related discrepancies on dual-channel arrays [138], quantile normalization for high-density oligonucleotide experiments [14]).

4. Log-ratios are calculated based on normalized intensity data. For each dual channel array, the log-ratio is calculated between its own hybridizations. When multiple probes target the same gene (technical replicates), to obtain one log-ratio per gene, the average is taken in case of a low number of replicates, whereas Median Polish is applied on the log-ratios if a large number of replicate probes are available to obtain a robust result (summarization).

   For single-channel arrays, the normalized intensities from technical replicates are first summarized to produce one intensity measurement per gene before calculating the log-ratios based on the contrast definitions (summarization). Median Polish is utilized when the number of replicates per gene exceeds 4, e.g. for Affymetrix; otherwise the average is taken.

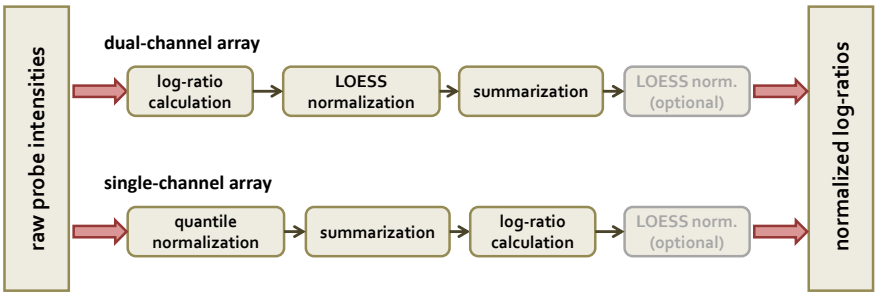


Figure 2.3: ***Default pipelines for compendium raw data homogenization***

5. After calculating log-ratios, the data quality is checked on a MA plot. Often, we still observed some non-linear intensity-dependent differences in the data, especially when Median Polish is utilized to handle the data generated by large number of replicates. This is then corrected through an extra non-linear normalization (e.g., loess fit). Similar discoveries and handling approaches have been described in the literature ([23, 134]).

The default data preprocessing pipelines for single channel array (Affymetrix, NimbleGen, Agilent one-color, etc) and dual-channel arrays are shown in Figure 2.3. Following each step, the data quality is checked visually. If there are issues with the result, extra steps and/or alternative methods and/or parameter settings are utilized to correct them. Generally, the homogenization is applied per experiment. Occasionally, some single-channel experiments conducted by the same lab share data generated on some arrays (the same access id in GEO), especially the reference ones probing wild type. In those special cases, although originated from multiple experiments, the contrasts sharing the same reference are homogenized together to achieve better consistency. After log-ratios are calculated for each experiment individually, they are integrated into one big compendium data matrix containing the data of all the processed experiments.

Multiple-chip platform data normalization For a multiple-chip platform, we first followed the same procedure to normalize data and calculate the log-ratio for each chip separately. Then an additional step is introduced, in which the log-ratios from multiple chips of the same platform are combined per contrast. Although the chips of a multiple-chip platform are designed to be complementary, sometimes there are a few genes that are measured on more than one chip generating multiple expression values per gene. To obtain a single gene expression value per contrast, the median is taken over the multiple chips to obtain one final value for each gene.

Data publishing

After an internal control, new compendia are released for public access. An existing compendium can be updated in two ways, an incremental revision and a new release. A *revision* adds extra experiments to the current version of the compendium providing more data. The new data are collected and processed independently, and directly integrated into the data of the current version of compendium after quality checking. Over time, as the understanding of the genome advances, there can be a genome revision for a species in which its gene annotations change. Recall that our compendium is a matrix of which rows correspond to genes, the change of gene annotation will change the number of rows this matrix has. In this case, a *release* is initialized creating a new version of the compendium incorporating the latest gene annotations. To create a release, the new gene annotations are first imported into the system, then the platform annotations are updated to these latest gene annotations, next all experiments available in the current version are homogenized again utilizing the updated platform annotations to generate the data matrix for a new version of the compendium.

2.2.3 COMMAND web system

As utilizing our methodology to create a compendium consists of multiple complex steps, a web system named COMMAND (COMpendium MANagement Desktop) (Figure 2.4) has been developed integrating the methodology with a web interface to facilitate compendium creation and maintenance. The system is composed of three components: the backend providing core functionalities for the creation of the compendium, the MySQL database to store the data collected from the online repositories and the content of created compendia, and the web service providing the front-end that interacts with users. The Apache server interfaces the communications between those components. The functionalities provided in the backend are computationally intensive and require no human interactions. These include but are not limited to compendium initialization, experiment information retrieval, raw data download, automatic data parsing, data homogenization, and publishing. The peripheral functionalities are implemented directly in the web system, such as, user control and management, new compendium initialization, etc. So are those functionalities that require human interactions, such as, various manual inspection functions, data annotation, etc. Additionally, data dependencies between various steps of the methodology are controlled through the web interface.

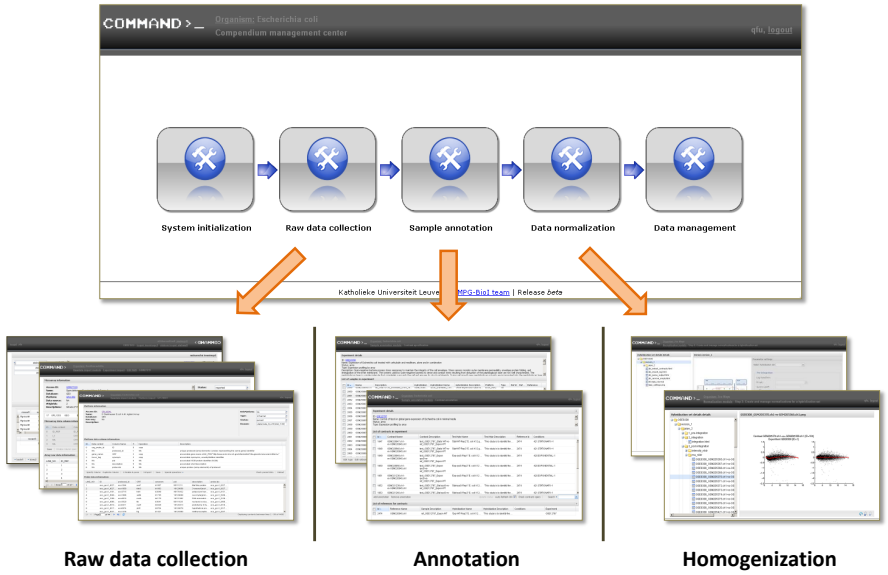


Figure 2.4: ***COMMAND compendium creation en curation system***

Availability The compendium creation system can be installed and configured to run on any LAMP server with Matlab runtime environment installed. Practically, a dedicated MySQL server can improves the performance of the system. The code is not publicly available but can be provided upon request.

2.3 Results and Discussion

Here, we present a methodology that consistently integrates high quality expression data of different experiments and platforms to create species-wise expression compendia accompanied by high quality manually curated annotations that facilitate both the global scale analysis and the targeted data exploration.

Similar efforts that directly integrate publicly available data across experiments do exist, e.g., M^{3D} [42] and Genevestigator [58]. Centering only on data generated by single platform (mostly Affymetrix), these efforts focus mainly on resolving the issue of inconsistently normalized data and on manual experiment metadata curation. The platform restriction simplifies data collection and

normalization tasks, but limits the amount of data can be integrated into the compendium. Our methodology does not impose such a limitation, instead tries to incorporate as much data as possible by integrating data across different platforms.

To achieve our goal, we faced the challenges related to data collection, data normalization, and integration. Compared to that of the single-platform compendium, the data collection is more complicated in two aspects. First, there is a lack of standard formats to report the raw expression values measured on various platforms. Especially for dual-channel platforms that probe expression values for two biological samples simultaneously, it is not trivial to identify and extract the desired data for each channel from data table of variant formats. Second, as two samples are hybridized on an array, the corresponding sample of the data generated in each channel must be correctly identified to guarantee that the proper experimental metadata information is used for annotation. To this end, a semi-automatic raw data parsing subsystem capable of handling data format heterogeneity and extrapolating correct data sample associations was developed. In the majority of the cases, experimental data available online are handled automatically by the system, avoiding time-consuming human intervention. When necessary, the extra information obtained from manual data inspection can be incorportated. The system enables processing a large amount of heterogeneous data efficiently.

Data generated from different types of platforms also complicates the data normalization and integration. In our system, a platform-specific normalization schema is utilized to handle particular issues associated with each type of platform, for example, loess to correct dye bias in dual-channel arrays, quantile normalization to reduce the variance among replicates for single-channel platforms. Different types of platforms produce different expression measurements. Single-channel platform generates absolute intensity value for each gene. Whereas, for dual-channel platform, log-ratio is generally preferred. As calculated between the intensities obtained from the same probe, it effectively removes undesired spot and array effects from the data. The log-ratio has been shown to improve consistency among results obtained across different microarray types [115], hence it is adopted as the type of the expression data for our compendium. To calculate log-ratios for all data, the concept of contrasts is introduced, in which one sample designated the test is compared with the other sample designated the reference to reveal differentially expressed genes. For a dual-channel array, a contrast is naturally defined between samples hybridized to it. For an experiment using single-channel arrays, one or more reference samples are manually chosen based on the platform used and the nature of the experiment. The log-ratio can then be effectively calculated for each contrast with a clear biological meaning representing the observed gene

expression changes induced by either the external perturbation(s) or the internal differences between samples. Furthermore, for experiments using single-channel platforms, it has been shown that the log-ratios calculated between samples from the same batch (experiment and platform) effectively remove undesired batch-effects [85].

Gene expression data are useful only when the accompanying sample information is available. We have also taken great care to provide an extensive formal contrast annotation and associated that with higher level condition ontology. As our compendium data are the log-ratios representing gene expression changes, our annotation focuses on specifying the differences between the pair of samples of a contrast. This is different from that of the single-platform compendium aiming to specify the complete information of each sample.

The methodology described here enables us to create a cross-platform expression compendium. Through integrating the data across different platforms, the methodology has two advantages compared to the existing single-platform one. First, it enables the creation of a compendium that incorporates much more data, hence providing a more complete expression landscape of a species. Second, it enables to create a sizable compendium for species which there is no dominant platform, such as Affymetrix. Utilizing this methodology, we successfully created three bacterial compendia for *Escherichia coli*, *Bacillus subtilis*, and *Salmonella enterica* serovar Typhimurium. The *E. coli* one is the largest currently available, and the last two are unique. The details of these bacterial compendia are described in chapter 3. Furthermore, the methodology has been adapted to construct compendia for eukaryotes. A proof-of-concept compendium for *Zea mays* (chapter 5) has been created, providing high data consistency by incorporating a complete re-annotation of the existing platforms based on the latest genome release and the extended annotations reflecting the complex structure and life style of a plant.

Chapter 3

COLOMBOS: access port for cross-platform bacterial expression compendia

3.1 Introduction

Over the last decade, the high-throughput omics technology led/driven by the microarray platform has revolutionized molecular biology research. With the unprecedented ability to globally probe gene expression, it was quickly adopted by most research labs generating colossal amount of data. With a few exceptions, most experiments designed to address particular biological question(s) are of *small-scale* covering a limited set of experimental conditions. Given the complexity of a living cell and its interaction with the environment, it has been shown that computational analysis of *large-scale* dataset across diverse experimental conditions and cell types is a powerful tool to reverse engineer regulatory networks [10, 39, 41], to study condition dependent behavior of such networks [40, 80], and to identify compound mode of action [10, 12, 51]. Although for many species, experiments publicly available in Gene Expression Omnibus (GEO) [9] or ArrayExpress [101] cover a broad range of the conditions, a direct integrated computational analysis of those data is not possible. Data are

This work has been published in K. Engelen*, **Q. Fu***, P. Meysman, A. Sanchez-Rodriguez, R. De Smet, K. Lemmens, A.C. Fierro, K. Marchal, COLOMBOS: access port for cross-platform bacterial expression compendia, *PLoS One*, volume 6, issue 7, 2011 (*These authors contributed equally to this work)

segregated per experiment due to the lack of uniformity in reporting expression data, the heterogeneity of the microarray platforms, and the incomplete and inconsistent meta information specifying experimental conditions.

Existing compendia alleviate the aforementioned data integration issues by either directly integrating data across experiments albeit limited by only those generated on a single-platform [42, 58] or combining results of individual experiments through meta-analysis to avoid direct data integration across experiments [36, 70, 99, 106]. Compendia created by meta-analysis are limited by the predefined functionalities provided by the system, hence lack of flexibility to incorporate new analysis. The compendia created by direct integration, however, retain actual expression values, hence the compendia can be readily analysed by existing and future methods. For eukaryotic model organisms, such a single-platform approach works well as the compendium based on Affymetrix chips can achieve a broad scope of experimental condition due to the platforms popularity. For prokaryotes, however, the single-platform constraint severely limits the scope of the compendium created due to the lack of such a dominant platform. For example, for model organisms such as *E. coli*, even the most popular platform 'Affymetrix GeneChip E. coli Genome 2.0 Array' is used in less than one third of the experiments. Consequently, a significant portion of data is missed out on when considering only one platform.

To have the advantage of direct integration, while not being limited to a single platform, we have devised a strategy that directly integrates expression data across platforms and experiments to create a compendium that provides an unprecedentedly broad coverage on experimental conditions. As such a cross-platform compendium benefits most the microbiology research community, we have applied our method to create compendia for three bacterial species: *Escherichia coli*, *Bacillus subtilis*, and *Salmonella enterica* serovar Typhimurium. Furthermore, to increase their usability for a large community of microbiologists, these compendia are being made available through COLOMBOS (COLlection Of Microarrays for Bacterial OrganismS). It is a web portal that provides easy access to the compendia and has an integrated suite of data tools for exploring, visualizing, and analysing the expression data. In this Chapter, we first briefly outline the compendium methodology, and then focus on the content of the bacterial compendia and the COLOMBOS access portal. Interested readers can find the detailed description about the methodology in Chapter 2.

3.2 Methods

3.2.1 Cross-platform expression compendium

The bacterial compendia are created following the methodology described in Chapter 2. It consists of three main steps designed specifically to remove the aforementioned hurdles that prevent direct data integration. First, the gene expression data and the corresponding experiment and platform information are extracted from GEO and ArrayExpress, removing the prevalent representation discrepancies. Next, a contrast is defined between two samples, whose gene expression levels are measured in the same experiment, one as reference and the other as test. The annotations are curated for each contrast by carefully analysing both the information stored in the online databases and the corresponding publication(s) (when available), specifying both the characteristic differences (e.g. the strain information) and the changes in the experimental condition between the test and reference samples. At last, the raw expression data are first normalized per experiment using dedicated procedures that respect the characteristics of the used microarray platform, and subsequently log ratios are calculated for each contrast between its pair of carefully chosen samples, representing gene expression variations caused by the sample difference, the experimental condition difference, or the combination of two. The log ratio calculation, capable of removing certain technical variations from the normalized data, inherently improves the data consistency across platforms and experiments [114, 115]. Following our methodology, the compendium created is, *de facto*, a matrix containing log-ratio expression values, in which each row correspond to one gene and each column one contrast.

3.2.2 COLOMBOS data analysis tools

COLOMBOS provides a suite of intuitive tools for exploring, visualizing, and analysing the expression data in the compendia. The interface is divided in two main sections: a 'Workspace panel' to the left and a 'Data analysis panel' to the right (Figure 3.1). The workspace panel is always visible: it contains the main control elements and shows an overview of the data (the 'workspace') the user is working with. The right hand data analysis panel is where querying of the database and visualization and analysis of the expression data takes place.

All steps and procedures in the COLOMBOS analysis tools act on what we call expression 'modules'. A module in COLOMBOS can be considered as a result of a single query to the database and is always a combination of a set of genes and a set of contrasts with corresponding expression values. Modules

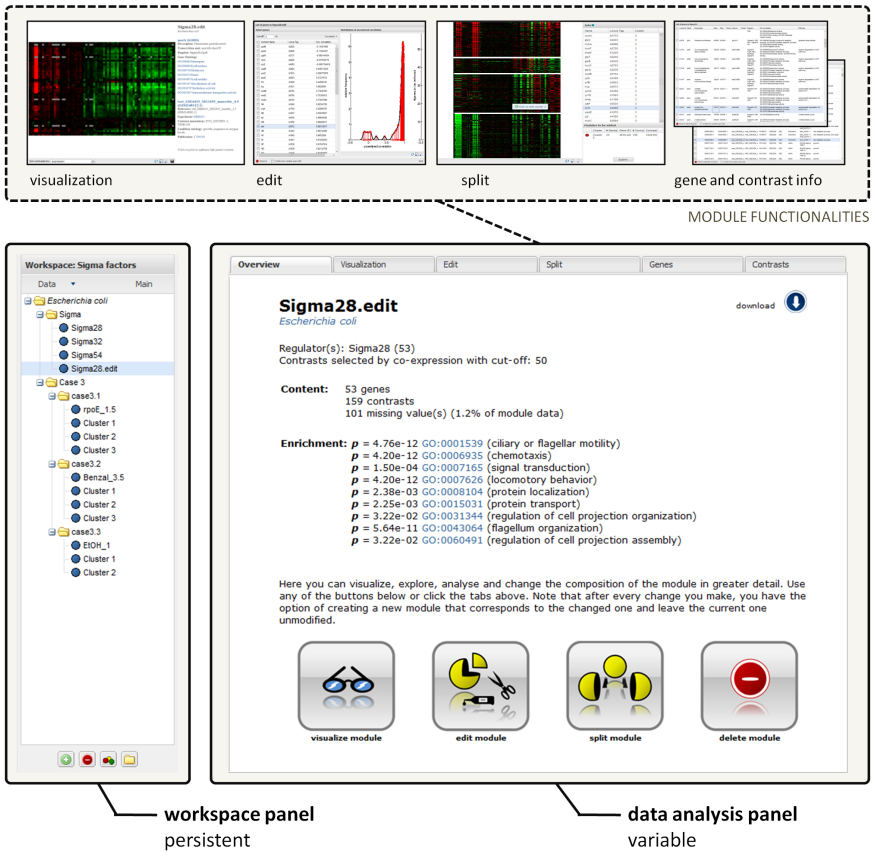


Figure 3.1: ***COLOMBOS data analysis components****. The bottom part shows the two main panels of the data analysis page. The left hand workspace panel is always visible, containing an overview of the modules and the main analysis controls. The content of the right hand data analysis panel depends on the actions of the user. In this case it shows the overview page for a module selected in the workspace. This overview page not only provides some general information on the selected module, but also serves as a guide for further examination and analysis steps. These are illustrated at the top part of the figure and include visualization, content editing (demonstrated is the removal of genes based on expression profile similarity), splitting the module based on expression values (shown here in the gene direction), and exploration of gene and contrast information.*

are dynamic in that at any time after creation their content can be altered by the user in various ways. In addition, multiple modules can be retained and organized in the workspace and can be analysed simultaneously. As the basic modus operandi, modules create a general framework through which various interesting, but conceptually different biological questions can be handled.

Three different options are given for creating a module: (1) by manually selecting only genes and have COLOMBOS automatically identify relevant condition contrasts, (2) by manually selecting only condition contrasts and have COLOMBOS automatically identify sets of co-expressed genes, (3) by explicitly selecting both genes and condition contrasts manually. Depending on the gene annotations that are available for the selected organism in the public databases that COLOMBOS integrates (see Table 3.1), the set of genes can be selected as anything from an operon or a regulon, to enzymes representing a metabolic pathway, or any custom list of genes of interest. Similarly, the module contrasts represent the biological conditions of interest and can also be retrieved in various ways, such as by experiment, by contrast annotation, or by condition ontology. When specifying only a set of genes, COLOMBOS will identify relevant condition contrasts based on the expression values of the selected genes in the compendium (user defined relevance cut-off that prioritizes both the magnitude as well as the consistency of the expression changes; see Appendix A for more details). Starting from only condition contrasts, COLOMBOS retrieves the most variable genes for the defined contrasts and (as an optional step) can identify clusters of co-expressed genes within this selection, which can be added as distinct modules.

Once a module is defined, it can be visualized in an interactive manner (with the option to export high-quality images), its expression values and contrast annotation can be downloaded, it can be split up in multiple modules in either the gene or contrast direction by clustering the expression profiles, or it can be further edited in gene and/or contrast composition by using available gene and contrast annotations or by analysis of the expression values in the compendium. These functionalities of the analysis tools are illustrated in Figure 3.1, showing the overview page for a single module. The module overview page gives some basic module information (such as the number of included genes and contrasts, the number of missing values, and a list of Gene Ontology enrichment scores) and serves as a helping guide to further analyse and visualize the module's composition.

When multiple modules have been created, they can also be explored and edited together. Any number of modules can be collectively visualized (to explore potential overlap), can be merged into a new module, and can be subtracted from one another in gene or contrast content. Visually exploring the module overlap, both in gene and contrast composition, can serve as an important guide

for deciding which modules may be grouped or subtracted.

Note that all of COLOMBOS' calculations, in both creating and editing modules, explicitly take into account the relative nature of the expression values by recognizing 0, implying no change, as the natural reference state of a log-ratio (details in Appendix A). Gene profile similarities are calculated by default as the uncentered Pearson correlation, which assumes that the sample means (i.e. the means of two gene expression profiles across a set of condition contrasts) are zero. Standard deviations of gene profiles are calculated in a similar way (as the root of the mean sum of squared log-ratios)

3.3 Results

3.3.1 Bacterial compendia

Currently COLOMBOS provides access to fully annotated public expression compendia for three bacterial model organisms: *Escherichia coli*, *Bacillus subtilis*, and *Salmonella enterica* serovar Typhimurium (see Table 3.1 for an overview of their respective content). These expression compendia are essentially organism-specific matrices of expression values derived from publicly available microarray experiments which are homogenized to make them comparable. The rows of a compendium matrix correspond to the known genes of the organism in question. Each column is a 'condition contrast' because it does not represent a single experimental condition, but in fact always represents the difference between a test and reference condition (the expression values themselves are calculated as expression log-ratios). Converting absolute measures of expression into expression changes is the principal means for rendering expression values comparable across platforms and experiments. Relative expression calculated intra-experiment/platform (i.e. between two conditions measured in the same microarray experiment using one platform) negates much of the platform and experiment specific variations that makes it otherwise impossible to reliably compare the absolute quantities reported in different experiments [115].

In order to be able to interpret and compare the expression log-ratios across an entire compendium, we have extensively annotated all contrasts using a set of formal hierarchically-structured condition properties (representing for instance mutations, compounds in the growth medium, treatments, and general growth conditions). This contrast annotation is done to structure the large amounts of potentially useful information that remain untapped due to the non-standardized condition descriptions in public databases. The annotation is complemented with a condition ontology that groups the condition properties

Table 3.1: An overview of the content of the three bacteria expression compendia

	Escherichia coli	*Bacillus subtilis*	*Salmonella enterica* **serovar Typhimurium**
Number of genes	4295	4105	4525
Number of contrasts	1429	259	717
source DB	GEO, AE	GEO	GEO
microarrays	1483	265	723
experiments	84	9	25
platforms	35	13	9
Missing values	6.1%	6.40%	3.90%
Condition properties	242	67	77
Condition ontology terms	56	24	23
External DBs			
pathway	EcoCyc	BioCyc	BioCyc
regulon	RegulonDB	DBTBS	
operon	EcoCyc	BioCyc	BioCyc
GO	UniProt GOA	UniProt GOA	UniProt GOA

under one or more ontology terms. It serves as a higher level organization, and provides a biologically more intuitive view of the condition contrast annotation by assigning properties of seemingly distinct categories to the same biological process. For example, in our *Escherichia coli* compendium the condition ontology term 'response to oxygen levels' includes condition properties that are linked to cellular processes that are dependent on oxygen availability, such as *fnr* mutations (a global oxygen responsive transcriptional regulator), NO_2 concentration (an electron transport decoupler), agitation of the growth medium, actual oxygen levels, etc. Apart from a thorough description of the represented biological conditions, we have incorporated several sources of information from main curated databases (UniProt GOA [18], EcoCyc [73], BioCyc [19], RegulonDB [50], and DBTBS [116]) into each of the microbial compendia. This includes additional data regarding gene function and genomic organization, metabolic pathways, and transcriptional regulation mechanisms. Both the condition annotation and additional gene information are integrated into the COLOMBOS data analysis tools in a functional manner to interactively browse and query the compendia (see Methods). If users so desire however, they can download the compendia in their entirety.

3.3.2 Case study - Fur regulatory targets

In the following case study we illustrate the benefits of exploiting the direct integration of expression values, as well as the ease with which one can make interesting biological discoveries using the COLOMBOS data analysis tools (see Methods for a detailed description of their functionalities). A straightforward application provided by COLOMBOS is the ability to find genes which show similar expression behavior with a starting set of genes for relevant condition contrasts. Since co-expression might infer co-regulation, we can use this approach to obtain a list of potential target genes that might also be regulated by the same transcription factor. In this example, we will use COLOMBOS to identify novel potential targets for the Fur transcription factor of *Escherichia coli*. Fur mostly regulates genes related to iron homeostasis and is strongly conserved across many Gram-negative and Gram-positive bacteria [22]. It has received a lot of interest in the past for its role in iron-limited conditions, such as those encountered by pathogenic strains in their hosts [100]. Fur has mostly been reported as a direct repressor of its target genes, but is considered a dual regulator: activation occurs indirectly by transcriptional repression of a small antisense RNA RhyB [89]. Fur has also been known to mediate combinatorial responses along with many other transcription factors [102, 145]. In the latest release of RegulonDB [50], Fur is described as having 98 target sites in 43 distinct promoters, with 28 of these promoters known to be subject to combinatorial regulation. The results of all data analysis steps discussed here are available in the case study data set accessible from the COLOMBOS home page at [26].

An initial set of 39 genes of the Fur regulon was constructed using the regulatory information integrated in COLOMBOS. Only genes known to be regulated by Fur alone, or by Fur in combination with the global regulators CRP, H-NS and/or FNR were selected. All other cases where known combinatorial regulation could occur were not included in the initial set because they might result in more complex, less homogeneous transcriptional responses. For similar considerations, if the activating sigma factor was known, only genes responsive to the household $\sigma 70$ were retained in the initial set. For this initial gene set the most relevant condition contrasts in the compendium were then selected, i.e. the contrasts where these genes showed the highest and most coherent response: a relevance cut-off (details in Appendix A) of 1 resulted in 97 contrasts. Not all of the retained genes show a similar expression profile for the retained contrasts however, which might be attributed to unknown active forms of combinatorial regulation or the dual regulatory function of Fur. Since we wanted to continue with a set of strongly co-expressed genes, COLOMBOS was used to further clean the initial gene set by removing genes that had a correlation smaller then 0.8 with the mean of the initial set for the selected contrasts. Next we used COLOMBOS to extend the remaining set of 30 genes with additional ones

that follow the same expression pattern for the selected contrasts (a correlation bigger than 0.8 was used as cut-off value), under the assumption that these constitute potential Fur targets. In this way, 19 extra genes were retrieved (Table 3.2), 7 of which were part of the Fur regulon but were not included in the initial set because they were known to be subject to regulation by additional transcription factors. The fact that these Fur-regulated genes were nevertheless retrieved might indicate that the additional combinatorial regulation was not active under the surveyed conditions.

Of the 12 novel genes, most showed a high likelihood of being Fur targets (Table 3.2). Six of these genes (*yqjH*, *ydiE*, *ybaN*, *yncE*, *yddB* and *ybiX*) were previously predicted to have a Fur target site in their transcription unit promoter by at least one of two independent studies [94, 100] (in case of *ybiX* as part of the proposed *fiu_ybiX* operon). Transcription of three of these (*ydiE*, *yncE* and *ybiX*) was moreover shown to be altered in a specific Fe^{2+}-Fur-dependent manner [90] and while little is known with regard to their function, the *ybiX* gene encodes a protein similar to an iron-regulated hydroxylase-encoding gene from *Pseudomonas aeruginosa*, further supporting a role for Fur in its transcriptional regulation. *pqqL* presents an interesting case: it encodes for a putative zinc peptidase and is situated directly downstream of the predicted Fur regulated *yddAB* operon in the chromosome. Using COLOMBOS to select the most relevant condition contrasts for the three genes *yddA*, *yddB*, and *pqqL* (see loadable case study data set available at COLOMBOS site) indeed shows that these genes are subject to tight co-expression, opening up the possibility of them being transcribed as a single transcription unit and putting *pqqL* under influence of the *yddA* promoter. The *feoC* gene is annotated as part of *feoABC* transcription unit as of the latest RegulonDB release (v6.8), which was not yet incorporated in COLOMBOS at the time of the analysis. This places it under the influence of the *feoA* promoter, which is a known Fur target. The *bfd* gene is clearly functionally related to Fur, being involved in iron storage and release, and has predicted binding sites in its promoter [22]. *bfd* is also the first gene in the *bfd_bfr* operon, *bfr* encoding an iron storage protein that is at the very least indirectly regulated by Fur as it has been shown that the expression of this gene is repressed by a small RNA RhyB, which in turn is repressed by Fur [89]. The complex Fur dependent regulation of *bfd_bfr* is also apparent by diverging expression responses for some of the selected contrasts. In the *E. coli* K12 strain, the gene *efeO* is part of an operon that has been disrupted due to a frame shift mutation. However, a Fur binding site was recently predicted in the *efeU* promoter [94] and it has been shown in the related *E. coli* Nissle 1917 strain that expression of *efeUOB* increases in response to iron-depleted conditions in a Fe^{2+}-Fur-dependent manner [55].

COLOMBOS also provides the functionality to retrieve anti-correlated genes,

Table 3.2: Finding potential novel Fur targets – a case study

Locus tag	**Name**	**Description**	**Operon**	**Known**	**COLOMBOS**	**Meta analysis**	**Evidence**
b1681	*sufD*	SufBCD Fe-S cluster scaffold	*sufABCDSE*	+		+	Fur, OxyR, IHF, lscR
b1683	*sufB*	SufBCD Fe-S cluster scaffold	*sufABCDSE*	+	+		Fur, OxyR, IHF, lscR
b2392	*mntH*	Manganese transport protein	*mntH*	+	+	+	Fur, MntR
b2673	*nrdH*	Glutaredoxin-like protein	*nrdHIEF*	+	+	+	Fur, NrdR
b2674	*nrdI*	Not annotated	*nrdHIEF*	+	+	+	Fur, NrdR
b2675	*nrdE*	Ribonucleoside-P_i reductase 2 α	*nrdHIEF*	+	+	+	Fur, NrdR
b2676	*nrdF*	Ribonucleoside-P_i reductase 2 β	*nrdHIEF*	+	+	+	Fur, NrdR
b4291	*fecA*	Fe3+ dicitrate transport protein	*fecABCDE*	+	+		Fur, CRP, PdhR
b0468	*ybaN*	Inner membrane protein	*ybaN*		+		Predicted
b0804	*ybiX*	PKHD-type hydroxylase	*ybiX*		+		Predicted; Fur dependent expression
b1018	*efeO*	UPF0409 protein	*efeUOB*		+		Predicted; functional in related strain
b1452	*yncE*	Uncharacterized protein	*yncE*		+	+	Fur dependent expression
b1494	*pqqL*	Probable zinc protease	*pqqL*		+		Potential operon *yddAB_pqqL*
b1495	*yddB*	Uncharacterized protein	*yddAB*		+		Predicted
b1705	*ydiE*	Not annotated	*ydiE*		+		Predicted; Fur dependent expression
b2211	*yojI*	ATP-binding ABC transporter	*yojI*		+		
b3070	*yqjH*	Uncharacterized protein	*yqjH*		+	+	Predicted
b3337	*bfd*	Bacterioferritin-associated ferredoxin	*bfd-bfr*		+		Indirect RhyB regulatio
b3410	*feoC*	Ferrous iron transport protein C	*feoABC*		+		TU *feoABC* with *feoA* known target
b4366	*bglJ*	Transcriptional activator protein	*yjjQ-bglJ*		+		

which can be interesting to investigate the potential of dual regulation (activation or repression by the same regulator). In the case of our Fur module, none of the anti-correlated genes pass the threshold of 20.8, but it is interesting to note that the second best ranked gene (correlation 20.74) is *ftnA*. This gene was not yet assigned as a Fur target in the RegulonDB release included in COLOMBOS, but it was recently shown that *ftnA* is transcriptionally activated by Fur directly (as opposed to indirectly through RhyB as is usually the case for Fur mediated activation) by reversal of H-NS silencing [97].

While the retrieval of already known Fur regulon genes combined with a set of likely targets confirms that a careful co-expression analysis can lead to the identification of novel targets, this does not imply that the direct integration of expression data itself, as in our compendia, provides any benefits. To illustrate the advantage of using cross-platform compendia, we repeated the analysis on a per experiment basis (a 'meta-analysis' of 7 experiments from which the 97 contrasts above were selected). Note that, to maximize the quality of the results of this meta-analysis, we did not use all contrasts within each experiment, but only the most relevant ones (selected with the same relevance cut-off as before), and that we ignored experiments with two contrasts or less. When extending the initial 30 genes with the same correlation cut-off of 0.8, the number of additional genes for each experiment ranges between 389 and 1385, the union adding up to a total of 3361. Most of these genes are false-positives with respect to being members of the Fur regulon: within single experiments generally only a limited number of similar conditions are surveyed and this increases the chance of finding genes with similar up and down regulation patterns but not sharing the exact same regulatory program. Trying to counter this effect by increasing the correlation cut-off does not necessarily yield better results, a cut-off of 0.9 resulting in the union containing 2135 additional genes, one of 0.95 in 1361 genes. Therefore we retained only the intersection, i.e. those genes that were added by each of the per experiment extensions with a correlation cut-off of 0.8. This intersection constituted 8 additional genes (a cut-off of 0.9 resulted in only 4 added genes, 0.95 resulted in none), 6 of them already known Fur targets, and only two uncharacterized genes representing potential novel targets. All of these were also retrieved by the COLOMBOS cross-platform analysis, with the exception of a single already known Fur target, *sufD*. However, another gene of the *sufABCDSE* operon was selected by the cross-platform analysis (*sufB*; all other genes of the operon showed correlations with the initial set of just under 0.8), retrieving the same promoter as a Fur target.

Table 3.3: Conceptual comparison of COLOMBOS with similar initiatives

	COLOMBOS	**M3D**	**GXA**	**GeneVestigator**
Database Content				
Expression data[1]	Cross-platform compendia	Single platform compendia (Affymetrix)	Experiment centered (ArrayExpress meta-analysis)	Single platform compendia (Affymetrix)
Organisms	Prokaryotes (3)	Prokaryotes (2) and a eukaryote	Eukaryotes (10)	Eukaryotes (9) and a prokaryote
Gene annotation	Incorporation of multiple species-specific DBs	Referal to SGD, BioCyc	EBI	None
Microarray annotation	Microarray annotation and condition ontology[2]	Microarray annotation	Microarray annotation and condition ontology[2]	Microarray annotation
Tools suite	Interactive visualization, expression analysis	Visualization, expression analysis	Interactive visualization, expression analysis	Interactive visualization, expression analysis
Functionalities				
Expression analysis[3]	Multiple queries	Single query	Single query	Single query (limited)
Query genes by ...	Gene IDs; functional or structural characteristics	Gene IDs	Gene/protein IDs	Gene IDs
Query arrays by ...	Experiment, annotation, or ontology	Experiment, annotation	Experiment, annotation, or ontology	Annotation
Download	Analysis results and/or entire compendia	Analysis results and/or entire compendia	Only experiments indirectly (through ArrayExpress)	Analysis results (limited)

1 Compendium: a data matrix (genes in rows, microarrays in columns) combining expression measurements from different experiments (an experiment being a set of microarrays submitted to the public DBs as such, implying that they were performed by the same lab and on the same technological platform). Single- vs. cross-platform: combining data from the same technological platform is relatively easy as the same preprocessing methodology can be employed; COLOMBOS is unique in combining data from different platforms using a specialized homogenization pipeline. Meta-analysis: expression data are not combined directly but experiments are analysed separately where after the results are compared.

2 The biological conditions measured on a microarray are described with a set of formal terms which are organized into a higher level ontology. Such an ontology facilitates querying for related experiments or conditions.

3 Single versus multiple queries: query results can be retained in the COLOMBOS user workspace where they can be organized and structured, into larger 'analysis projects'. This allows for integrative across-query analysis where relations between single query results can be explored, e.g. by combining or differentiating single query results.

3.4 Conclusions and future directions

In this work we aimed at closing the gap towards an encompassing expression resource for prokaryotic organisms and facilitate the use of information in publicly available microarray experiments for a large community of microbiologists. We have created fully annotated cross-platform expression compendia for three bacterial model organisms: namely *Escherichia coli*, *Bacillus subtilis*, and *Salmonella enterica* serovar Typhimurium. These compendia can be accessed through a web portal called COLOMBOS which also provides a suite of integrated analysis and visualization tools. To our knowledge, COLOMBOS is unique in offering compendia for *B. subtilis* and *S. Typhimurium*, and its *E. coli* compendium is the largest currently available. To maximally exploit the available expression data, several aspects of both compendia construction, as well as design and implementation of the analysis tools, are exclusive to COLOMBOS (see Table 3.3 for a conceptual comparison with similar initiatives). Most notably, the compendia were created by directly integrating expression measurements from different experiments and microarray platforms. The reputed low reproducibility between microarray experiments and platforms [8, 125] (although more promising findings have also been reported [76, 115, 113]) is not a legitimate argument for not combining them: short of an objective basis to dismiss certain measurements, a lack of agreement between two experiments does not render either invalid and might in fact be a strong motivation to integrate them. In our previous research directly combining expression data from different sources proved a valuable asset for reconstructing transcriptional networks [40, 80, 142], and here we wanted to take the principle of direct cross-platform integration to a higher level by generating large scale expression compendia with a broad applicability for biological discovery. Directly integrating expression data enables one to simultaneously assess multiple diverse conditions, relevant to the biological problem of interest and ensures a finer-grained view of condition dependent transcription responses that can lead to higher quality predictions, such as in the case study above for extending the known regulon of a transcription factor.

We have also taken great care to provide an extensive formal condition contrast annotation and associated higher level condition ontology for all compendia. Microarray experiments that are committed to a public database, such as ArrayExpress or GEO, are required to comply to the MIAME standards [15, 16]. And while much effort has been taken to standardize the description of the experimental protocols used in a microarray experiment, there are no specifications of the format in which the surveyed biological conditions should be presented. The resulting cryptic, non-standardized condition descriptions in public databases do not enable computational comparison and automatic

organizing of experiments which our annotation does. Another feat in which COLOMBOS is unique: this condition annotation is functionally integrated in the data analysis tools allowing the user to interactively browse and query the compendia, not only for specific arrays or experiments, but also for specific experimental conditions and biological processes. In a similar fashion, information from main curated microbial databases is also integrated to interactively browse and query the compendia for specific genes, pathways, transcriptional regulation mechanisms, and more.

Downloadable versions of the entire annotated compendia, as well as the data analysis tools, are available at COLOMBOS web portal [26]. In a half-yearly fashion new revisions of the compendia, updated with additional experiments, will be made available. We also plan to increase the current scope of organisms by adding new compendia for other bacterial species using a flexible framework for creating and updating cross-platform compendia which is currently in development. The data analysis tools incorporated in COLOMBOS will continue to be developed to offer users enhanced tools for analysing and visualizing the compendia's expression data.

Chapter 4

Directed module detection in a large-scale expression compendium

4.1 Introduction

Omics based approaches are increasingly being used to uncover underlying mechanisms of bacterial behaviour [46]. The obligate deposit of high throughput experiments in public databases upon publication has tremendously increased the amount of publicly available experiments. Mining these data sources helps in gaining a global condition dependent view on bacterial gene regulation [40, 80], in expanding the current knowledge on transcriptional interactions with novel reliable predictions [39, 41], and in comparing transcriptional networks across species [4, 46]. It also offers molecular biologists the possibility to see their own dedicated analysis in light of what is already available. Inferring transcriptional networks from these public data usually requires complex normalization procedures [42] and computationally intensive algorithms. To enhance the usability of these tools, they are often wrapped in a web service.

This work has been published as a Book Chapter: **Q. Fu**, K. Lemmens, I. Thijs, P. Meysman, A. Sanchez-Rodriguez, H. Sun, A.C. Fierro, K. Engelen, K. Marchal. Directed module detection in a large-scale expression compendium, Van Helden, J.; Toussaint, A.; Thieffry, D (eds.), *Methods in Molecular Biology - Bacterial Molecular Networks (MMB)*, 2012

In this chapter, we will illustrate how compendia of public expression data can be used to identify condition dependent coexpression modules, in which a particular gene of interest is involved ('directed' module detection), by means of two web services: COLOMBOS [38] and DISTILLER [80] (the latter in combination with the visualization tool ViTraM [123]). Figure 4.1 illustrates the difference between both approaches. COLOMBOS only relies on expression data to retrieve condition dependent modules, implying that there is a functional relationship between the module genes, but not necessarily that they are regulated by the same (set of) transcription factor(s). DISTILLER, in contrast, incorporates additional motif data and the constraint that module genes should share the same regulatory program, implying that the module's condition dependent coexpression can be directly linked to transcriptional coregulation.

To clearly demonstrate the functionalities of these web services, the gene *sodA* is used as a case study. This gene encodes for a protein with superoxide dismutase activity, which reduces harmful free radicals of oxygen formed during normal metabolic cell processes [44, 68, 126]. It is known to be regulated by several regulators, such as, Fur, SoxS, and MarA. As such its expression is coupled to different biological conditions: multiple antibiotic resistances (MarA), superoxide resistance (SoxS) and the intracellular iron pool (Fur). By applying different analysis approaches (COLOMBOS and DISTILLER), we will show how to identify *sodA* containing modules, i.e. genes that behave similarly to *sodA* in a condition dependent way, from large-scale expression compendia.

4.2 Materials

4.2.1 Cross platform expression compendium

An expression compendium is essentially an organism-specific matrix of expression values derived from publicly available microarray experiments which are homogenized to make them comparable. The rows of the matrix correspond to the known genes of the organism in question. Each column is a *contrast* defined as a comparison between two different biological samples, one acting as a test and the other as a reference. The expression values are calculated as expression log-ratios representing gene expression changes induced by the difference between samples. Relative expression calculated intra-experiment/platform (i.e. between two conditions measured in the same microarray experiment using one platform) negates much of the platform and experiment specific variations that makes it impossible to reliably compare the absolute quantities reported in different experiments [115]. The extensive annotations are manually curated for each contrast, specifying which aspect(s) (experimental conditions) has/have been

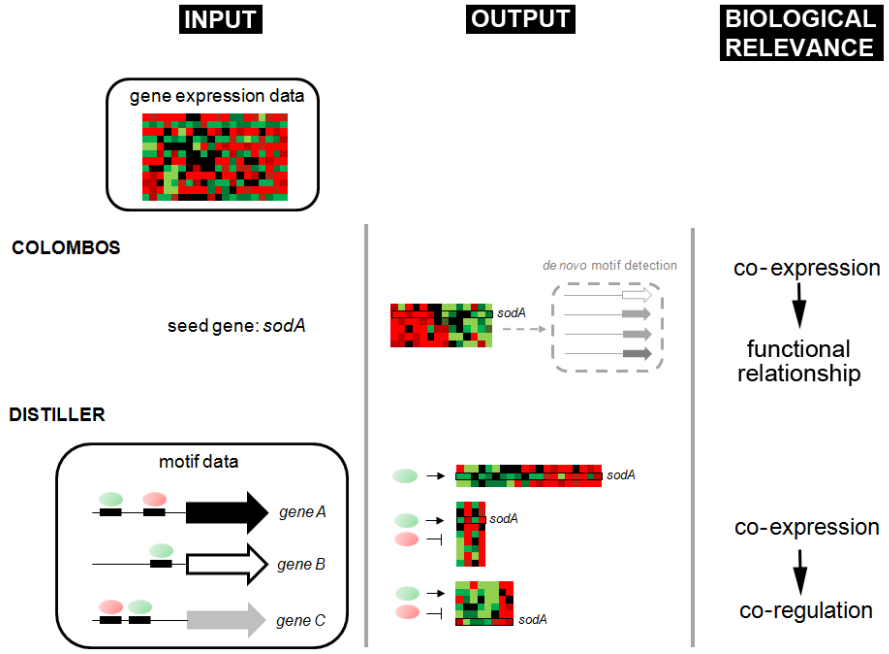


Figure 4.1: ***Workflow of different methods for the directed module detection in expression compendium.*** *Coexpression modules are detected by COLOMBOS starting from a user provided seed gene (e.g., sodA). DISTILLER searches for coexpression modules in a global fashion (information on seed genes is not a prerequisite) and is constrained by regulator-to-gene assignments. Coexpression patterns retrieved by COLOMBOS point toward a functional relationship among the module genes but do not necessarily imply coregulation. To gain insights in the regulatory program of those genes, de novo motif detection tools could be applied to analyze their promoter regions. Inherent to the regulator-to-gene constraints implemented in DISTILLER, there is a direct link between coexpression and coregulation for the genes in the modules that it retrieves.*

changed between the test and reference samples. The methodology utilized to create such a compendium is explained in detail in Chapter 2.

The *Escherichia coli* compendium used here contains 1429 condition contrasts obtained from 1747 microarrays of 84 experiments across 35 platforms, covering the expression profiles of 4295 genes under various conditions annotated by 242 different condition properties grouped under 56 condition ontology terms. Furthermore, several sources of gene information from main curated databases are integrated into the compendium as well, such as, the pathway and operon information from EcoCyc [73], the Gene Ontology annotation from UniProt

GOA [18], and the regulon information from RegulonDB [50]. (see Table 3.1)

4.2.2 Coexpression modules

Genes that have the same regulatory program (here defined as having the same transcription factor binding motifs in their promoter regions) behave similarly: their expression responds in a similar manner to certain alterations in the organism's intra- or extracellular environment. This is called *coexpression*. The evidence of coexpression points to a functional relationship between coexpressed genes, and might sometimes be an indication of shared regulatory programs. Such a functional relatedness is only enforced when coexpression occurs across multiple different conditions. Since each value in our data set does not represent an absolute measure of expression, but rather a relative one associated to a contrast (test versus reference), coexpression in this case translates to genes showing log ratios that are significantly different from zero for at least one condition contrast, and for each of the involved contrasts show coherent changes in the same direction (either up or down). The methods, COLOMBOS and DISTILLER, demonstrated in this chapter employ different computations to score this coexpression, but essentially look for the same phenomenon.

Genes are only coexpressed under certain condition contrasts. This combined set of genes and condition contrasts, where the coexpression pattern appears, is what we refer to as a *module*. Extra constraints can be used to tune the module concept for specific purposes. For example, extra requirements could be added that genes in the module should have (known) motif(s) in common.

4.2.3 COLOMBOS

COLOMBOS, an acronym for *COLections Of Micrcroarrays for Bacterial OrganismS*, is a web service [26, 93] (Figure 4.2) for interactively exploring, querying, analyzing and visualizing data from cross platform expression compendia through an intuitive interface. Currently it provides the possibility to query expression compendia based on the annotation of their contrasts and/or genes. This service can be used, for instance, to search for genes being coexpressed with one or more genes of interest under a pre-specified set of condition contrasts, or to search for the condition contrasts in an expression compendium under which a pre-specified gene set is coexpressed. As such, it can identify coexpression modules based on user input. In COLOMBOS, modules can be visualized as interactive heatmaps, showing the annotation of genes and condition contrasts as obtained from public databases and corresponding literature. More details of COLOMBOS are discussed in Chapter 3.

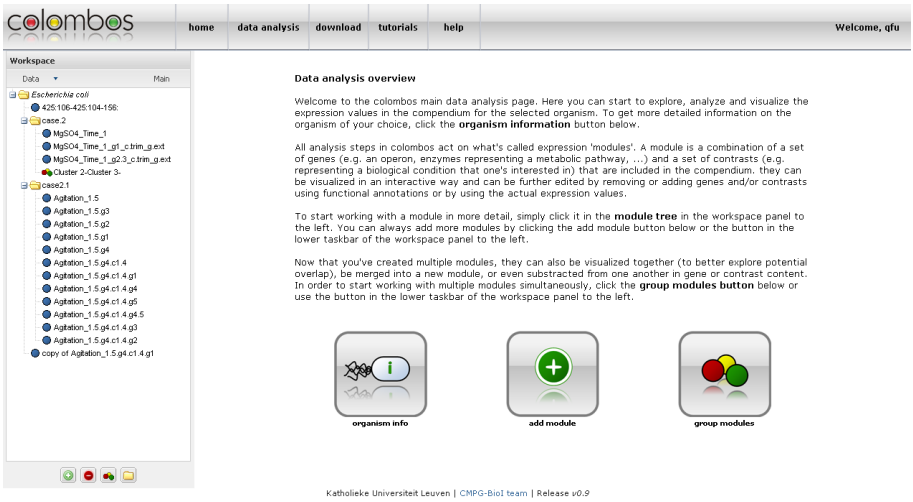


Figure 4.2: ***COLOMBOS web service interface.*** *The interface is composed of three parts: the website navigation header at the top, the workspace (showing a module tree) at the left-hand side, and the operating space (here showing the 'Data analysis overview' panel) filling the center.*

Currently COLOMBOS (v2.0) runs on three browsers: Firefox, Google Chrome, and Opera. Be sure that you are using the latest version of the browser to have the best compatibilities. Please check the website for updated information.

COLOMBOS expression data format

To use COLOMBOS requires no input file, as the expression compendium is an integral part of it. It does provide a tab delimited ASCII file format for expression data export, containing information on the user created module consisting of a particular set of genes and the corresponding set of contrasts under which these genes are coexpressed. There are two sections in this file. The first one describes the condition information of the module. It specifies for each contrast the name and description of its test and reference sample, the corresponding experiment identifier, the online database from which the experiment was retrieved, the microarray platform used to measure the sample hybridizations, and finally, the annotated condition changes between the test and the reference sample. The second section contains three parts: the corresponding contrast identifiers in the first row; the gene information (gene locus tag, gene name, and internal gene identifier) in the first three columns; and the expression log-ratios of the compendium. Starting from the 4^{th} column, each column in

the file corresponds to one condition contrast. Each of the rows represents one gene and its expression values for different condition contrasts.

4.2.4 DISTILLER

DISTILLER (*Data Integration System To Identify Links in Expression Regulation*) is a data integration framework that searches for condition dependent transcriptional modules by combining expression data with information on the interaction between a regulator and its corresponding target genes, for instance based on motif screening. A module is here defined as a set of genes coexpressed under a sufficiently large number of conditions and sharing the motif instances of the same regulator(s). The expression data used by DISTILLER could be either absolute expression values or log-ratios.

DISTILLER searches the modules in a global fashion over the entire dataset through three steps: identification of candidate modules, module filtering, and module extension. Built on advanced itemset mining approaches, it first exhaustively enumerates all possible valid *closed* modules (see next Section) as candidates. The candidates are partially redundant as they might share the same genes or condition contrasts. Therefore, in the filtering step, modules are prioritized by calculating an interest score for each of them. The score takes into account the probability that a module of the same size can be found by chance in the input dataset, whereas, at the same time, penalizes the overlaps between modules. The resulting top ranking modules are statistically significant yet distinctive. These initial modules obtained under stringent thresholds are called '*seeds*', in which the identified interactions between a transcription factor and its target genes are highly reliable (high precision). However, some true interactions might be missed (lower sensitivity). Hence, in the last module extension step, the modules retained after filtering are further extended with additional genes by applying relaxed thresholds on minimal coexpression level and motif score. Note that it is not required to execute all three steps. The result of each step can be analyzed and visualized independently.

The DISTILLER web service [34] (Figure 4.3) is designed to leverage the difficulty of applying the algorithm by providing an easy-to-use and consistent interface that links all three steps together. The interface is separated into two sections vertically. The information panel at the top provides a brief explanation of the site's functionality. The project panel at the bottom is where the user interacts with the site. It is further divided horizontally into three sections. At left, the 'Project Information' panel provides information of the current project and the links to access the input, output, and log files. The 'Operation' panel at the center is where a user runs the DISTILLER algorithm. It has

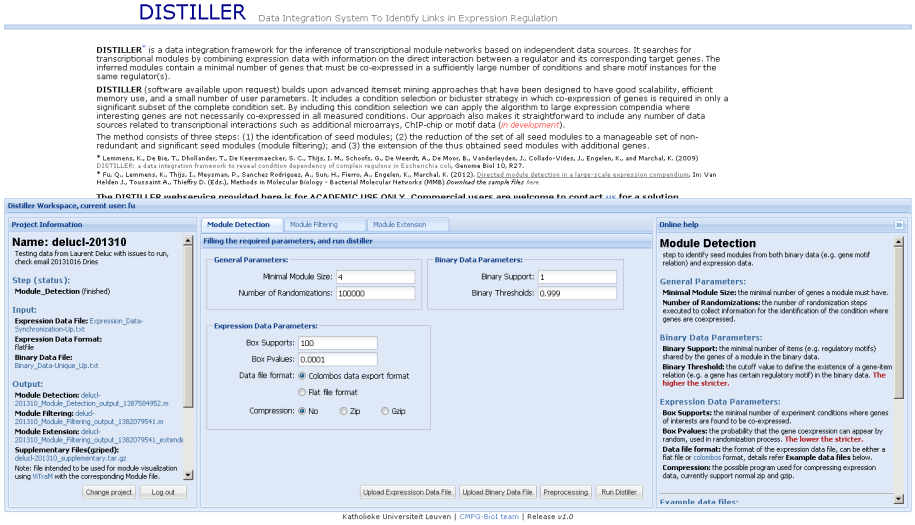


Figure 4.3: ***DISTILLER web service interface (at module detection step).***

three tabs, each of which provides access to one step of the algorithm, namely, module detection, module filtering, and module extension. The explanation for the parameter settings and the input data file format (if required) of the corresponding step selected in the 'Operation' panel are provided in the 'Online help' panel at right for your reference. The website works properly on Firefox and Google Chrome. Efforts are undertaken to make it compatible with other browsers as well. Please check the website for updated information.

For this case study, the data consist of the same compendium as the one accessed through the COLOMBOS website. Hence, the columns of the expression matrix, across which DISTILLER defines coexpression, correspond to the condition contrasts as defined in the compendium (see Section 4.2.1).

Closed module A module is considered *closed* if it cannot be further extended by any other gene without reducing the number of motifs shared by all of its genes. To search such a candidate, the algorithm [80] starts with the smallest modules that contain only one gene, and gradually extends them by merging with other modules. A candidate is found when any extension violates the conditions implied by 'closed'. Only closed modules that contain the minimal number of motifs and condition contrasts are considered valid.

DISTILLER input file format

DISTILLER requires two input files: an expression data file and a motif data file. In the case study described here, the expression data file contains the expression log-ratios for each gene under each contrast. Two possible input formats are allowed for the expression data file. The first one is a matrix based flat file format with header, in which, the first row contains the identifiers of each contrast, each subsequent row corresponds to a gene, each column represents a contrast. Gene name is used to specify a gene in the first column of each row. The second format is that of the COLOMBOS expression data export file (see Section 4.2.3). The system further supports the use of two most common compression formats to handle the large data set, namely, Zip[1] and Gzip.

The motif data file contains the results of a motif screening analysis [57]. Each motif screening score indicates the probability that a motif instance is present in the promoter region of a gene, with 1 as the maximum score and 0 the minimum. The data is specified in a text file containing a data matrix with header, in which the header are motif names, the row corresponds to gene, and the column represents the motifs screened. Similarly to the expression data, except for the header, each row starts with gene name specifying the corresponding gene.

Additionally, users can provide an operon information file that will be used during the module extension step. It contains two items in each row, the gene that belongs to an operon, and an identifier of that operon.

In DISTILLER, the different data files are coupled in the gene direction, so the user should make sure that the same gene identifiers are used in each file, although the order of the genes need not be the same.

DISTILLER output file format

All three steps of DISTILLER result in the same output format: a '.m' file. It contains the information of the identified modules, with each module described in a section, separated by an empty line. Each section contains four data fields. Given a section specifying the information for a module M, the field 'Significances' holds the p-values that were assigned to M by the itemset mining algorithm [141] utilized by DISTILLER. The field 'items' contains the information of the genes that belong to M, each of which is represented by

[1] In case of zip file, the system allows only one file in the compressed file that is the input expression data file. An error is generated when this condition is violated.

its row number[2] in the expression data input file. The field 'boxtidset1' specifies the contrasts of M, in which the genes in M are coexpressed, each of which is represented by its column number[3] in the expression data input file. Finally, the field 'tidset1' contains the motifs of M shared by its genes, specified by their column numbers in the motif data input file. Each module is represented by a unique number indicated between the '{}' brackets after the name of each corresponding data field.

Besides the main output file, two supplementary files required to use the visualization software ViTraM are also provided. They are bundled in a compressed file that can be downloaded from the website. The file whose name starts with 'expdata_' contains the preprocessed input expression data, and the other starting with 'binary_' contains the preprocessed input motif data. The file formats of these two files are the same as the corresponding input files described in the previous section.

Output notification email Since each step of DISTILLER may take hours or even days to finish, the result is sent to the user by email. This notification email is of a standard format, with subject 'DISTILLER process result notification email'. On the first line, the information about the finished process and the corresponding user project is presented, followed by the link to the result file, then the link to the supplementary file bundle required for visualization.

4.2.5 ViTraM

To analyze DISTILLER output, we use ViTraM (*VIsualization of TRAnscriptional Modules*) to visualize overlapping transcriptional coexpression modules together with the motif information in an interactive way. Here, we will only briefly discuss this tool. For more details we refer to [123].

ViTraM 2.0 is used for this case study (download at [130]). Written in JAVA, it can run on any platform with JAVA support. On Windows ViTraM comes with two options: vitram_Windows_512M.bat or vitram_Windows_2G.bat. The former has a minimum memory requirement of 1Gb RAM. And we advise to use a computer with at least 3Gb RAM to run vitram_Windows_2G.bat.

[2]When counting the row number of a given gene, there are certain caveats. First, the header should not be counted. Second, the count should start from 0. Hence the first row is of row number 0 instead of 1.

[3]To count the column number referred in the '.m' file, the first column containing gene information should not be counted, and it starts at 0.

ViTraM input file format

ViTraM requires an XML file and an expression file as input. The XML file contains the information of the transcriptional modules that will be visualized: the genes and conditions composing the modules and additional information such as motifs that are assigned to modules by DISTILLER. Due to the complexity of the XML format, an extra tool called XMLCreator is available at the ViTraM website to automatically generate the XML file from the DISTILLER output file (see Section 4.2.4). Conveniently, the tool also generates a smaller expression data file containing only the relevant expression values from those genes and condition contrasts existing in the result modules obtained by DISTILLER. Moreover, the tool can also incorporate extra information into XML file, for example, the motif data contained in the binary supplementary file mentioned in Section 4.2.4. For the purpose of this study, we will not discuss this XML file format in detail. Interested users can find more information in [123].

4.2.6 Sample files

Four sample files are provided as example input files for this case study. Users can download them at this address [109] or from the DISTILLER website [34] (follow the sample file link of the second reference).

The file expdata_COLOMBOS_module_information.txt is an example of the exported module data file of COLOMBOS, which, as one of the input formats of the expression data for DISTILLER, contains both the gene and condition contrast information, and the log ratio expression data. The file expdata_DISTILLER_Expression.txt is an example of the other expression data format accepted by DISTILLER. The file binary_DISTILLER_Motif.txt is the motif data containing motif scores for each gene. And finally, the file operon_DISTILLER_OperonGenes.txt is the example operon information file.

Furthermore, three example output files, corresponding to the output generated after each step of DISTILLER, are provided. DISTILLER_Output.m is the output of the seed module identification step, DISTILLER_FilteringOutput.m is the output of the module filtering step, and DISTILLER_outputExtended.m is the output of the module extension step. Files expdata_DISTILLER_supple.txt and binary_DISTILLER_Motif_supple.txt are provided as the supplementary files used for visualizing modules contained in those sample output files.

Finally, ViTraM_Modules.XML is provided as the XML file describing the modules, and ViTraM_Expression.txt is provided as the expression data file used by ViTraM.

4.3 Methods

We will describe the steps required to use two web services, COLOMBOS and DISTILLER, to identify coexpression modules around a (set of) gene(s) of interest (query genes). This will be illustrated by applying them to the query gene *sodA* as a case study. The analysis flow based on COLOMBOS shows how it can be used to search for genes that are coexpressed with a (set of) query gene(s) under the human guidance. Alternatively, by combining expression data with motif data, the DISTILLER approach aims to detect coexpressed and coregulated gene sets globally in a single run in an unsupervised way.

4.3.1 Identifying coexpression modules using COLOMBOS

To search for coexpressed genes using COLOMBOS, we will first create an initial expression module by specifying a set of genes of interest (query genes), and use COLOMBOS to extract the most relevant contrasts, where the chosen genes are differentially expressed. Next, the module is extended with genes that are coexpressed with the initial gene set under the selected of contrasts. Here, we explain this workflow starting with a single query gene, *sodA*.

Create the initial module

Step 1 Go to COLOMBOS website [26], and create the initial module based on the known gene set of interest and the biological conditions that are known to affect these genes when changed.

Step 1.1 To start the analysis, click on 'data analysis' in the title bar (Figure 4.2) to bring up the data operation interface in the center part of the website. It is separated horizontally into two parts. At the left hand side, there is a 'workspace' information panel that lists all the modules created in a tree structure. This panel is always visible when residing on the data analysis page. Since no module has been created, it is currently empty. The other part of the page, referred as the operating space in this text, is where the visualization and analysis of the expression data takes place and is currently occupied by the 'Data analysis overview' panel (hereafter referred to as 'overview panel'). This panel serves as a intuitive guide showing the available analysis options depending on the current workspace state along with some brief explanations.

Since we just started a new session, there is only one button 'select organism' available. Click it, and select the organism '*Escherichia coli*', which this case study focuses on. Notice now this species appears in the workspace panel as the root of the tree.

Step 1.2 After the species is picked, the content of the overview panel is changed accordingly. The button 'select organism' is replaced by two other buttons 'organism

info' and 'add module'. To create our initial module for *sodA*, we click the 'add module' button to bring up the 'Add module overview' panel. In this panel, there are three buttons 'select gene only', 'select contrasts only', and 'select gene and contrast'. Each corresponds to a different method for creating a module. Here, we utilize the first option to manually specify the query gene *sodA* and let the system automatically identify the most relevant contrasts for it based on expression data. Clicking 'select gene only' brings up the gene selection panel at center. Four available options are available: 'By gene name/locus tag', 'By transcriptional regulation', 'By pathway', and 'By transcription unit'. The first option allows us specify gene manually. Click the yellow box surrounding the text 'By gene name/locus tag', fill in '*sodA*' in the text box appeared (one gene per row), and then click the 'Done' button to proceed.

Step 1.3 After specifying the genes, a 'ranked contrast selection' panel (Figure 4.4) appears in the operating space, which allows the user to select contrasts ranked by a score that prioritizing those show the highest magnitude of change and most coherent coexpression for the selected set of genes (see Appendix A.1). The panel is separated into two parts horizontally. At the left, the contrasts are ranked according to the score from highest till the lowest. A cutoff value can be specified in the box above the list to select only those with higher scores. The figures to the right show, from left to right, a density plot showing the distribution of the scores across the whole compendium and a plot showing the number of contrasts that will be added if a given cutoff is selected. There is no absolute guideline to choose the cutoff. Several try-outs may be needed to identify the optimal one.
For our initial module, we chose cutoff value 3. Specify it, and click 'Ok' button at

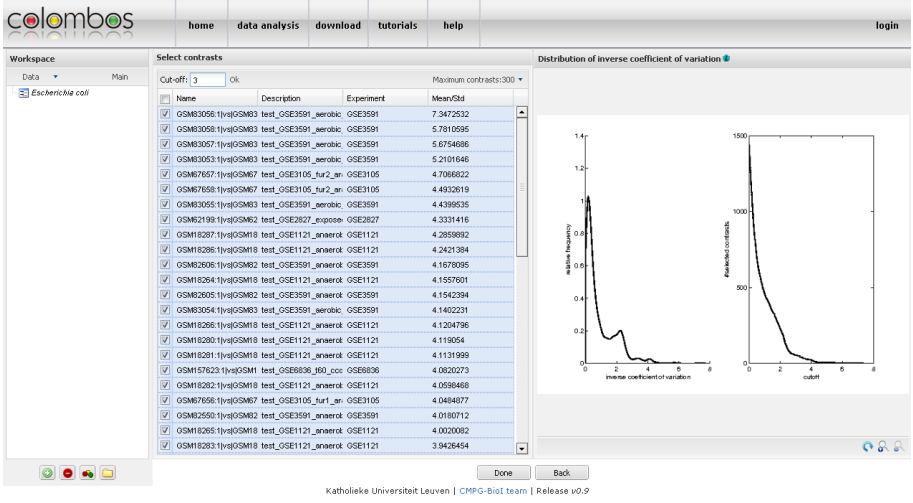


Figure 4.4: ***The 'ranked contrast selection' interface to choose contrasts based on the specified module genes (sodA in this particular case).***

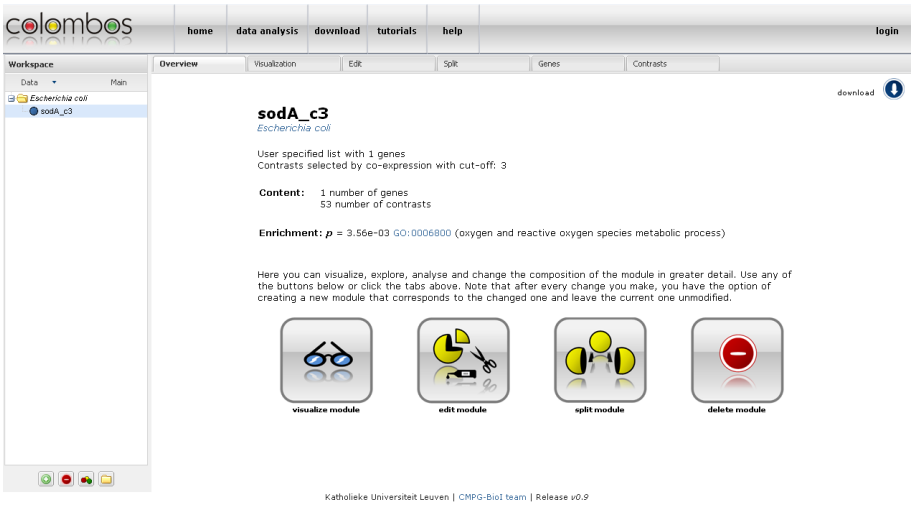


Figure 4.5: ***The overview tab of the module 'sodA_c3'.***

the side. The contrast list is then filtered, and the contrasts retained in the list are automatically selected to be added to the module. Click 'Done' button at the bottom, and fill in 'sodA_c3' as the module name to create our initial module. The newly created module appears in the module tree in the workspace as a node directly under the root (presenting the selected organism *Escherichia coli*).

Inspect the existing modules

Step 2 One can review the information of a module through various options.

Step 2.1 Click the module you want to check in the workspace to show its overview tab (Figure 4.5) in the center. The tab shows on top a brief summary of the module, including, name, description, number of genes and contrasts it contains, as well a list of significantly enriched Gene Ontology (GO) terms of class biological process (p-values < 0.1, details see Appendix A.4) for the module genes. At the bottom, there are four buttons to *visualize*, *edit*, *split*, or *delete* the module. The function of *visualization* and *delete* is straightforward. The *edit* function allows user to modify either the genes or the contrasts of the module. Instead, the *split* function breaks the current module into several new ones by separating either its genes or its contrasts into distinctive groups and creating a module for each selected group.

Step 2.2 Each module can be visually presented as a heatmap. In it, each square represents one gene's expression log ratio for a certain condition contrast. A red color indicates over-expression while a green one shows under-expression. The intensity of the color corresponds to the magnitude of the expression change. The brighter the

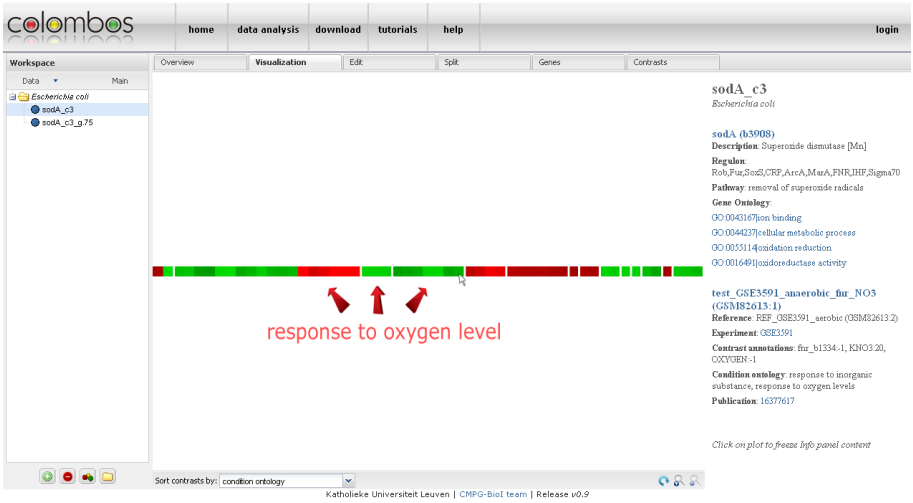


Figure 4.6: ***The visualization tab of module 'sodA_c3' showing the heatmap.*** *In this heatmap, the contrasts are grouped by their ontology annotations. The groups belonging to the ontology term 'response to oxygen level' are marked out.*

color, the higher the absolute ratio value. To view the heatmap, click 'Visualization' button in overview tab to switch to heatmap tab. The heatmap is shown at the center with an information panel at the right. Hovering over the heatmap, the information of the gene and the contrast under the cursor will be shown in the information panel. Various options exist to group the contrasts in the heatmap according to their experiments, condition properties, and the associated condition ontology terms. They can be selected from the dropdown box located at the lower left corner of the tab. The heatmap of our module sorted on condition ontology is shown in Figure 4.6.

Checking module 'sodA_c3', it shows that the top 53 contrasts, where *sodA* is most differentially expressed, have been selected. Among those, 34 contrasts belong to the condition ontology term 'response to oxygen levels'. It includes condition properties that are linked to cellular processes dependent on oxygen availability, such as *fnr* mutations (a global oxygen responsive transcriptional regulator), NO_2 (electron transport decoupler), agitation of the growth medium, actual oxygen levels, etc. This is expected as the function of *sodA* is linked to processes related to oxygen availability. COLOMBOS indeed successfully identifies them as the most relevant contrasts of *sodA*. There are also 17 contrasts belonging to the ontology term 'growth', with three in common with the aforementioned 34 contrasts. The term 'growth' is very general, grouping conditions that trigger various biological processes simultaneously at a global cellular scale. Hence it is not unexpected to find *sodA* differentially expressed

under these contrasts.

Extend the genes of a module

Step 3 Next, we will extend the module with additional genes. Click 'edit module' button to go to the 'Edit' tab, where three options are available: edit the name/description, the genes, or the contrasts. Click 'edit genes' to modify the module genes. A gene modification panel appears with all options to edit genes. The options in green boxes indicate ways to add genes to the module, whereas those in red boxes indicate ways to remove genes. For this case study, we will use the last option 'Add new genes based on expression' to add additional genes, whose expression patterns are most (anti-)correlated to the expression pattern of the module. Click it to bring up a 'ranked gene selection' panel for gene selection.
This panel (Figure 4.7) functions in a similar way to the 'ranked contrast selection' panel explained in Step 1.3. On the left side, a gene list ranked by the degree of the correlation of each gene's expression profile with the mean profile of the module genes under the selected contrasts belonging to this module. Different types of correlation scores are available to rank the genes. Interested users can refer to Appendix A.2 for the detailed explanation of the different options. Here, the default one, which calculates the *Uncentered Pearson Correlation*, is utilized. The higher the score, the more similar the profiles are. Similarly, a cutoff value can be specified to select only

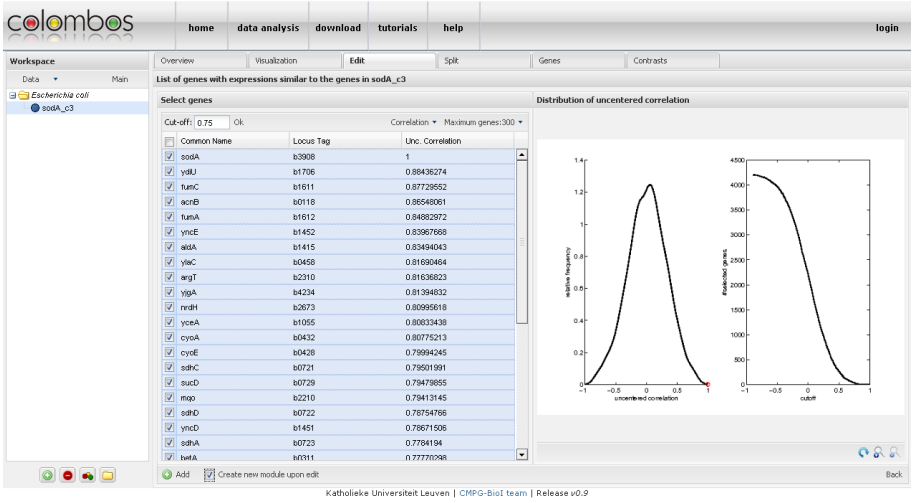


Figure 4.7: ***The 'ranked gene selection' interface to extend module 'sodA_c3' with the most relevant genes.*** *The red circle indicates sodA in the module. As it's expression profile is perfectly correlated with itself (the correlation score of 1), the circle is located at point (1, 0).*

those genes with higher scores. Also in the panel, the figures to the right show, from left to right, a density plot displaying the distribution of the correlation scores across all genes in the compendium and a plot showing the number of genes that will be added if a given cutoff is selected. The circles connected by a red vertical line in the density plot indicate the gene(s) that already exists in the current module and its correlation score.

To extend our module, we will use a cutoff value of 0.75. The option 'Create new module upon edit' (located below the gene list) is checked, so that a new module is created with the extended gene set and the original one is kept unchanged. When ready, click the 'Add' button at the left, and specify 'sodA_c3_g.75' (in the pop-up window that appears) to name the new module. A new node representing newly created module is added below the original module 'sodA_c3' in the workspace.

The resulting module now contains 35 genes with an expression profile similar to that of *sodA* under the selected contrasts (see heatmap, Figure 4.8). Amongst the genes in the module are *cyoA*, *cyoB*, *cyoC*, *cyoE*, subunits of the cytochrome b terminal oxidase complex involved in aerobic respiration [28], *sdhA*, *sdhC*, *sdhD*, encoding the succinate dehydrogenase (SdhCDAB) active during Krebs cycle catalyzing the oxidation of succinate to fumarate under aerobic conditions [135], and *sucA*, *sucB*, *sucC*, *sucD* involved in generating succinyl-CoA, one of the reactants in the Krebs cycle [17]. Considering the functions of these gene products (enriched GO terms in Table 4.2), it is not at all unexpected that their expression levels are influenced by oxygen availability. Moreover, many genes in module 'sodA_c3_g.75' also share common regulators with *sodA*, such as ArcA, Fur, FNR, CRP, etc. Hence, *sodA* (encoding a superoxide dismutase activity) being coexpressed with the genes involved in aerobic respiration might be essential to protect the cell against oxidative stress.

Summary

The biological relevance between genes and contrasts of module 'sodA_c3_g.75' clearly shows the strength of the simple methods implemented in COLOMBOS. Although the coexpression of module genes does not necessarily imply coregulation of its genes, the promoter regions of these genes could be investigated further using motif detection methods to discover common regulators (see Figure 4.1). This task could be achieved either by screening these regions with Position Specific Scoring Matrices (PSSMs) of previously characterized motifs, or by using *de novo* motif detection methods [119, 127]. Moreover, some genes in the module are unannotated. This shows that the COLOMBOS web service can also serve as a tool to help biologists identify interesting research candidates.

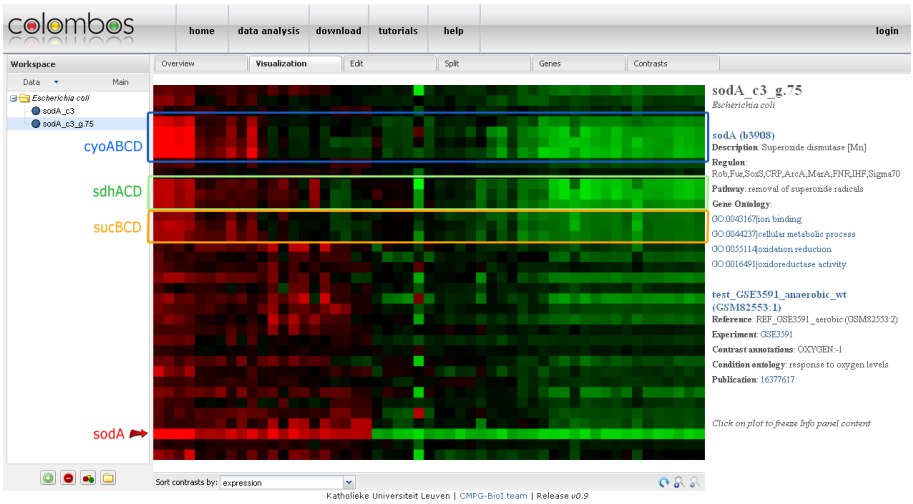


Figure 4.8: ***The heatmap of AU2 module 'sodA_c3_g.75'.*** *The row of expression data that corresponds to sodA is marked by the arrow. The expression data of the gene groups that are prominently coexpressed with sodA, namely cyoABCD, sdhACD and sucBCD, are marked by the named boxes.*

4.3.2 Identifying transcriptional modules with DISTILLER

DISTILLER [80] is a data integration framework that automatically identifies condition dependent transcriptional modules by combining expression data with information on the interaction between a regulator and its corresponding target genes (here through motif data). Here, we will first show how to setup the related parameters and run each of the three steps of DISTILLER through our web service [34]. Then after transcriptional modules are obtained, they are visualized together with the motif information in an interactive way using software ViTraM [123]. At last, we briefly discuss those resulting modules containing *sodA*.

Part1, Running DISTILLER algorithm

DISTILLER automatically identifies all regulatory modules that meet certain criteria from the global expression and motif data of a species. The target specific ones can then be filtered out from the resulting modules. In this case study, we will use DISTILLER to identify those modules containing *sodA*. Consequently, we can simplify the motif data by removing motifs irrelevant to

the query genes. This greatly reduces the running time of DISTILLER. After this simplification, only 7 motifs (ArcA, Rob, SoxS, Fur, IHF, FNR, MarA, and CRP) remain.

Create a user profile and a project DISTILLER algorithm is computationally expensive, as it explores the complete search space. Depending on the specified parameters and the dataset size, each step could take hours or days to finish. Consequently, a registration is required, which creates an account with minimally user's email information. This provides great flexibility, as users can simply setup the system to run a step of the algorithm and leave the site. When the process is finished, the user is notified by email (details in Section 4.2.4). He can then log in and select the corresponding project to continue.

Step 1 Go to the DISTILLER web service [34]. A new user can be created in three steps, click the 'New User' button, complete the required information, and click the 'Create User' button. The user's email address is required for sending notification emails (see Section 4.2.4).

Step 2 After log in, the project selection interface will appear, where one can either select an existing project to continue, or create a new one by clicking 'Create New Project' and giving a name and a brief description for it. Here, we create a new project named 'Ecoli_sodA'.

Step 3 After choosing or creating a project, the DISTILLER panel will appear where the user can run the three main steps of DISTILLER, each of which has its own separate tab. For a new project, only the 'Module Detection' tab is accessible (Figure 4.3), while the 'Module Filtering' and 'Module Extension' tabs are disabled.

Identifying seed modules

Step 4 For a new project, users first need to upload the required expression data file and motif data file (see Section 4.2.4) to be analyzed by the DISTILLER algorithm, and preprocess them to check their format and integrity. As the system supports expression data file in different formats and compression types, it is the user's responsibility to specify the correct options for the 'Compression' and 'Data file format' items in the 'Expression Data Parameters' section. Here, the sample files, expdata_DISTILLER_Expression.txt and binary_DISTILLER_Motif.txt, are used. Upload them using data file specific upload buttons, specify 'Flat file format' as 'Data file format' and 'No' for 'Compression', then click the 'Preprocessing' button to start the process on the server. As the process is time consuming, the user will be notified by email when it is done.

Step 5 After receiving the notification email, the user can go back to their project at DISTILLER site. The interface is the same as seen in Step 4. Only now the button 'Run Distiller' is unlocked. Three groups of parameters need to be specified before

starting the process: the binary data parameters for the motif data (see Step 5.1), the expression data parameters (see Step 5.2), and the general ones (see Step 5.3).

Step 5.1 Binary data parameters: 1) 'Binary Support' specifies the minimal number of motifs that the genes in a module should have in common; 2) 'Binary Thresholds' is the minimal score a motif instance in the promoter region of a gene should have to be considered as present. As DISTILLER only accepts binary motif data as input, whereas most motif detection algorithms generate probabilities scoring the certainty of motif gene relations, the specified threshold is applied on the motif data to convert the *continuous values* into *binary ones*. For this case study, we choose 1 for the Binary Support and 0.999 as the Binary Threshold. Hence the genes belonging to a module must share at least 1 motif and a gene is considered to have a specific motif if the corresponding value in the motif data is equal to or higher than 0.999.

Step 5.2 Expression data parameters: 1) 'Box Support' specifies the minimal number of condition contrasts under which the genes in a module should be coexpressed; 2) 'Box P-value' is used to generate the threshold bandwidth sequence that is needed to test whether a module passes the constraints on the expression data, i.e. whether the genes in the module are sufficiently coexpressed within the selected contrasts [80]. Here, 100 are chosen for the 'Box Support' and 0.0001 for the 'Box P-values'.

Step 5.3 General parameters: 1) 'Minimal Module Size' specifies the minimal number of genes that should be in a module; 2) 'Number of Randomizations' specifies the number of random modules that will be generated for computing a threshold bandwidth sequence for the condition contrast selection. Here, we specify 4 and 10000 for these two parameters respectively. These require that the algorithm generates 10000 random modules consisting of 4 genes to compute the threshold bandwidth sequence.

Step 6 When all parameters are specified, click 'Run Distiller'. DISTILLER will now enumerate all initial modules, each of which contains at least 4 genes coexpressed in more than 100 contrasts and sharing at least one motif. Note these contrasts are selected based on comparing the ordered contrasts bandwidth sequences of given genes with the threshold bandwidth sequence (see [80] for details). When finished, the user will receive a notification email (see Section 4.2.4) containing the link to the output file with the identified modules, and the link to the supplementary file bundle that is required to visualize modules using ViTraM.

Filter the raw DISTILLER output After obtaining the initial modules, the user can go back to the saved project at the DISTILLER site. Now, the tabs 'Module Filtering' and the tab 'Module Extending' are enabled. Note that the user can directly proceed to the module extending step, but due to the discussion in Section 4.2.4, performing a filtering step before extending the obtained modules is highly recommended.

Step 7 Click the 'Filtering' tab, specify the number of modules to be filtered out in this step, then click on 'Filter Result' to proceed. In our case study, the top 20

modules ranked by their scores computed by DISTILLER are selected (see Section 4.2.4). Similarly, the user will be notified of the result of the process by email.

Extend DISTILLER seed modules After obtaining the filtered results, the user can now proceed with the 'Module Extension' step. In this step, extra genes are incorporated into a module if they comply with the relaxed criteria.

Step 8 Two parameters need to be specified to run this step. Candidate genes have to comply with both parameters in order to be included in the extended module. The only exception is made for the genes belong to an *operon* (see Step 9).

Step 8.1 'Extended Motif Threshold': this is the same type of parameter as the 'Binary Thresholds' specified in Step 5.1, but less stringent to allow a gene having more present motifs. As a result, more genes can satisfy the 'Binary Support' parameter. Here, we use 0.95 for this parameter (as compared to 0.999 for the 'Binary Thresholds').

Step 8.2 'Correlation Percentage': a correlation threshold to select candidate genes for module extension. Given a 'Correlation Percentage' of p, only those genes whose expression profile correlations, calculated based on the mean expression profile of a module, higher than p, are considered as the candidates. Here we choose p as 0.95.

Step 9 Optionally, the genes of a module can be further extended with *operon information* if available. Candidate genes belonging to an operon, whose first gene is present in a seed module, only need to satisfy the criterion for the expression profile ('Correlation Percentage') to be included into the module. The operon information (i.e. which genes belong to which operon) is included in the analysis by uploading a file containing the corresponding data (see Section 4.2.4).

Step 10 After specifying the parameters and optionally uploading an operon data file, the user can click 'Extend Modules' to run the process. A notification email is sent when the process is finished.

After receiving the final output file of DISTILLER, the user can then specifically select the modules that contain his or her query genes, in our case the gene *sodA*, to continue.

Part2, Visualization of DISTILLER modules with ViTraM

The software ViTraM is used to analyse modules discovered by DISTILLER. However, the DISTILLER output files need to be converted into ViTraM compatible format first. Then, ViTraM can be utilized to visualize the modules.

DISTILLER is non-deterministic due to the randomization process utilized by the algorithm to select contrasts. Thus, it is possible that when following exactly the same way as outlined here to run DISTILLER, the recovered

modules are different from those obtained in this example (as provided in 'DISTILLER_outputExtended.m' sample file, see Section 4.2.6). Use the corresponding sample file to proceed with exactly the same results.

Prepare ViTraM readable files

Step 1 Download the XMLCreator from the ViTraM website[130] and unzip the file. Under linux, run the XMLCreator by typing the command 'java -jar -Xms256m -Xmx512m XMLCreator.jar' in a terminal. Windows users can click file 'XMLCreator_Windows_512M.bat' to run it.

Step 2 In the pop-up window, choose 'DISTILLER' as module detection tool and click on 'OK'. The main interface appears where user can specify the input and output files. Two required input files are the DISTILLER '.m' output data file and the 'Expression Data' file (see Section 4.2.5). In addition, other files providing extra information can be visualized along with the modules. These include but are not limited to, the motif information of each gene, the genes' functional annotations, or the experimental factors and/or sample characters of each contrast, etc. Finally, the names of the output files need to be specified. Two output files will be generated: the module XML file and the corresponding expression data file containing the expression values for those genes and contrasts presented in any of the modules.

Here, we demonstrate this step using the corresponding sample files, the 'DISTILLER_outputExtended.m' and the 'expdata_DISTILLER_supple.txt'. Additionally, the file 'binary_DISTILLER_Motif_supple.txt' providing extra motif information for each module is specified as the 'Motif Data'. We then run XMLCreator to generate the two output files, 'ViTraM_Modules.XML' and 'ViTraM_Expression.txt'. To run XMLCreator on the DISTILLER output, please always use the supplement files specified in the notification email that accompanies the '.m' file (see Section 4.2.4).

Run ViTraM From the ViTraM website [130], the user can download ViTraM v2.0 and unzip the file. Note, a registration is required by providing us some basic information, and the user must accept our software license.

Step 3 To run ViTraM under windows, click on 'ViTraM_Windows_512M.bat' or 'ViTraM_Windows_2G.bat' provided. The latter can handle larger datasets, but does require more than 2Gb physical memory in the computer. To run ViTraM under linux or mac, call 'ViTraM_Linux&Mac.sh' file.

Step 4 Load 'ViTraM_Modules.XML' and 'ViTraM_Expression.txt' generated in Step 2 by operations 'Open Module XML File' and 'Open Expression File' in the 'Input' panel respectively. After the data are loaded, they can be visualized by 'Load All Modules' in the 'Module Selection' panel. This extra loading step enables visualizing a subset of modules, where directly visualizing all the modules in a large dataset generally causes memory issues. In such a case, users can use the 'Filter' panel first

to select only the relevant modules, then visualize them. This significantly reduces the loading time to visualize modules.

Step 5 Subsequently click on 'View Modules' in the 'Module Display' panel to show the modules in the main window. In this window, the genes are listed at the left and the conditions at the top, and each module is represented as one or a group of boxes with its id shown at the top left corner of each box and a distinctive color for each module.

Step 6 Click 'View Gene Properties' in 'Gene Properties Display' panel, the extra properties of the gene can be visualized together with the modules in an extra window located at the left side of the 'Module Display' window. The motif information provided in Step 2 will now appear here. The color of each cubic represents the value of the motif score explained in Section 4.2.5. The red color corresponds to the high value and the green the low one. If a score is higher than a threshold, a cross will appear in the corresponding cubic.

Step 7 By clicking on 'Favorite Genes' in the 'Filter' panel at the right hand side, it is possible to display only those modules containing a particular gene. The pop-up window shows on the left a list of all genes. Select the gene 'b3908' (*sodA*), and click the arrow button '→' to add it to the favorite gene list at the right. Then click 'OK'. Only modules containing *sodA* will still be in display.

Step 8 To optimize the visualization of modules that overlap, click on 'Automatic ordering' in the 'Module ordering' panel and subsequently 'Run Overlap Index' to layout the currently displayed modules in the most optimal way (for details on the algorithms that identify the optimal display of overlapping modules, we refer to [123]).

Step 9 Click on 'View Modules' and subsequently 'Refresh Modules' in the 'Overview & Heatmap Display'. Then click on 'Adding Heatmap'. The expression values of the genes for the condition contrasts in the currently displayed modules will be shown by means of a heatmap.

Analyzing resulting DISTILLER modules containing *sodA*

In this case study, we used DISTILLER to search for modules of coexpressed genes sharing at least one motif instances for the same regulator. Resulting modules including the motif instances shared among genes within each module are listed in Table 4.1. Since we are interested in *sodA*, transcriptional modules 2, 3, 8, and 12 containing this gene were selected and visualized using ViTraM (Figure 4.9). The numbering of the modules corresponds to their ranks in the filtered results. Note that as a global method, DISTILLER identifies all possible regulatory modules in the dataset (not only those with *sodA*). So the numbering of the modules containing *sodA* is not consecutive. The lower the number, the more significant a module is.

Table 4.1: Overview of the 20 transcriptional modules identified by DISTILLER

ID	Motifs	# Genes	# Conds	Genes in *sodA* modules
1	Fur	22	97	
2	MarA, SoxS	4	77	*fpr*, *poxB*, ***sodA***, *zwf*
3	ArcA, CRP	7	79	*acnA*, *acnB*, *aldA*, *gltA*, *osmY*, *sdhC*, ***sodA***
4	FNR, IHF	4	226	
5	ArcA, FNR	4	123	
6	CRP	308	107	
7	CRP, FNR	4	87	
8	Fur, CRP	4	146	*cyoA*, *nupC*, *sdhC*, ***sodA***
9	SoxS	4	181	
10	Fur	18	86	
11	MarA	4	206	
12	CRP, FNR	5	90	*aldA*, *cyoA*, *malP*, *pdhR*, ***sodA***
13	IHF	8	87	
14	IHF	53	123	
15	FNR	38	102	
16	CRP	76	125	
17	ArcA, CRP	4	148	
18	SoxS	4	95	
19	Fur	4	100	
20	Fur	22	76	

Module 2 contains the genes regulated by both SoxS and MarA. The identified contrasts in this module belong to the ontology terms 'carbohydrate metabolic process', 'growth', and 'response to oxidative stress'. SoxS and MarA are known

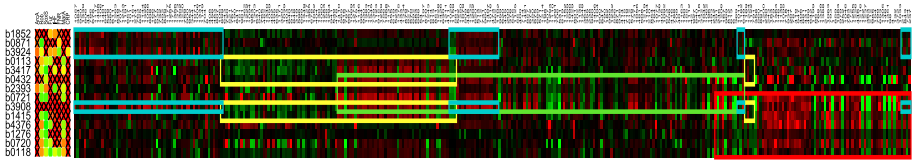


Figure 4.9: ***DISTILLER sodA related modules visualized in ViTraM.*** *The figure shows the four modules containing sodA. Each module is indicated with one or more squares and expression values are indicated by heatmap. Module 2 is indicated by blue lines, module 3 by red line, module 8 by green lines, and module 12 by yellow lines.*

Table 4.2: GO enrichment of sodA modules[1]

Module	P-value	GO ID	GO Term Description
COLOMBOS			
sodA_c3_g.75	2.41e-12	GO:0006091	generation of precursor metabolites and energy
	9.23e-02	GO:0006800	oxygen and reactive oxygen species metabolic process
	3.17e-02	GO:0006869	lipid transport
	2.74e-07	GO:0022900	electron transport chain
	4.22e-02	GO:0042592	homeostatic process
	1.85e-02	GO:0045454	cell redox homeostasis
DISTILLER			
module2	7.07e-02	GO:0005996	monosaccharide metabolic process
	7.11e-03	GO:0006800	oxygen and reactive oxygen species metabolic process
module3	3.11e-05	GO:0006091	generation of precursor metabolites and energy
	1.77e-02	GO:0006800	oxygen and reactive oxygen species metabolic process
	7.11e-02	GO:0044262	cellular carbohydrate metabolic process
module8	4.09e-03	GO:0006091	generation of precursor metabolites and energy
	1.06e-02	GO:0006800	oxygen and reactive oxygen species metabolic process
	4.09e-03	GO:0022900	electron transport chain
module12	1.42e-02	GO:0006800	oxygen and reactive oxygen species metabolic process

1 Enriched GO terms are ordered by GO access ID.

to participate in the removal of superoxide and nitric oxide and protection from organic solvents [128]. One can expect that the contrasts belonging to 'response to oxidative stress' to be suitable for inducing their expression and therefore the expression of genes regulated by them. We already highlighted the link between the removal of superoxide species and metabolic pathways. It explains why the condition contrasts under 'carbohydrate metabolic process' are found in this module. As discussed after Step 3 of Section 4.3.1, to observe condition contrasts belonging to 'growth' is not unreasonable in this case. In addition to *sodA*, the

other genes found in this module are *fpr* (Ferredoxin-NADP reductase), *poxB* (Pyruvate dehydrogenase), and *zwf* (Glucose-6-phosphate 1-dehydrogenase). The enriched GO terms of this gene set (Table 4.2) are highly consistent with the observed condition ontology terms.

In module 3, genes are regulated by both ArcA and CRP. Its condition contrasts are highly related to the ontology terms 'growth' and 'response to oxygen levels'. CRP is a global regulator involved in the degradation of any non-glucose carbon sources and also an antagonist of catabolite repression. On the other hand, ArcA participates specifically in a signal transduction system sensing particular aerobic and anaerobic growth conditions [27]. The functions of the genes found in this module (*acnA*, *acnB*, *aldA*, *gltA*, *osmY*, *sdhC*, and *sodA*) are an intersection between global metabolic pathways and specific processes responding to oxygen level changes (see Table 4.2).

Genes in module 8 are regulated by both CRP and Fur. As previously mentioned, CRP is a global regulator that facilitates bacterial fitness in function of the availability of different carbon sources. Fur is a sensor of intercellular iron concentration, and also participates in the response to reactive nitrogen species [96]. The contrasts of this module mainly involve 'carbohydrate metabolic process', 'growth', 'detoxification of nitrogen compound', 'lactose catabolic process'. In addition to *sodA*, the genes found in this module are *cyoA* (Ubiquinol oxidase subunit 2), *nupC* (Nucleoside permease nupC), and *sdhC* (Succinate dehydrogenase cytochrome b_{556} subunit). Altogether the condition contrasts and the genes selected in this module reflect the linkage between basic metabolism changes induced by external C-source availability (e.g. glucose concentration) and intra-cellular energy production through the electron transfer chain (see also enriched GO terms in Table 4.2).

In module 12, genes are regulated by CRP and FNR, both master-regulators. FNR activates genes involved in anaerobic metabolism and represses genes involved in aerobic metabolism. The enriched GO term (only 'oxygen and reactive oxygen species metabolic process') for this module's gene set reflect this. The selected contrasts of this module constitute a very broad set of ontology terms, including 'carbohydrate metabolic process', 'growth', 'DNA repair', 'SOS response', 'lactose catabolic process', and 'detoxification of nitrogen compound'. This is in line with the global regulation exerted by CRP and FNR.

4.4 Discussion and Conclusion

The COLOMBOS web service provides a very straightforward and intuitive way to analyze and explore large-scale expression compendia. It is intended

for specific queries based on the user's pre-knowledge on specific genes or conditions of interest. It is a deterministic approach, which means that when executing exactly the same operations, COLOMBOS will generate exactly the same result. Furthermore, the analysis methods implemented in COLOMBOS are fast, generally taking only seconds or less to generate results for each step.

DISTILLER, on the other hand, is a semi-automatic method. Except for a limited set of parameters, it requires very little user involvement. Guided by both motif information and expression data, the method tries to compose a global regulatory network that covers the whole input dataset (see Table 4.1). It is non-deterministic due to the randomization involved in generating a threshold bandwidth sequence for the condition selection (see Step 8 in Section 4.3.2).

In this case study, we have used COLOMBOS and DISTILLER to gain insights in the functional processes, in which *sodA* is involved, and the regulatory programs that coordinate them. This resulted in one module identified by COLOMBOS and four modules by DISTILLER. As discussed in previous sections, each module does reflect relevant biological processes, in which genes including *sodA* participate. This does not necessarily require the modules to overlap, as each might represent different processes involving different genes. When these two approaches are used to address the same biological question as was done here, the distinct features of each approach lead to different but meaningful and thus complementary results.

The module 'sodA_c3_g.75' identified by COLOMBOS contains several genes involved in various biological processes (see Figure 4.6 and the discussion at the end of Section 4.3.1) with very high and coherent expression values under most contrasts (in both cases where up-regulation or down-regulation occurred). In contrast, when compared to module 'sodA_c3_g.75', each of the DISTILLER modules contains in general much less genes, which show less extreme expression levels indicating a weaker coexpression signal compared to 'sodA_c3_g.75'.

When evaluating the overlap between modules found by these two methods, module 3 and 'sodA_c3_g.75' share four genes (*aldA*, *acnB*, *sdhC*, *sodA*) and 19 contrasts ('response to oxygen level', 'growth'). These contrasts represent approximately 25% and 35% of all condition contrasts of each module respectively. Module 8 and 'sodA_c3_g.75' have three genes (*cyoA*, *sdhC*, *sodA*) and eight condition contrasts ('response to oxygen level') in common, which represent approximately 6% and 15% of all condition contrasts in each module. This kind of similarity can be expected, since we applied both methods to answer the same biological question based on the same expression data.

On the other hand, no overlap is observed between module 2 and 'sodA_c3_g.75'. Module 12 shares three genes (*aldA*, *cyoA*, *sodA*) with 'sodA_c3_g.75' though,

but the two modules have no contrasts in common. Upon closer inspection of the expression data, genes in DISTILLER module 2 are coexpressed, but the expression value of *sodA* is not very high for these contrasts, when compared with those of 'sodA_c3_g.75'. Hence those contrasts will not be easily picked up by COLOMBOS in Step 2.3 of Section 4.3.1 (unless a much lower selection threshold is used). Furthermore, genes appearing only in the DISTILLER modules show a different expression behavior from that of *sodA* under those contrasts selected by COLOMBOS. As a result, when extending genes of a module based on expression profiles (see Step 4 of Section 4.3.1), those genes will be ranked as less relevant. Consequently, they will hardly be considered as prime candidates for module extension.

In summary, COLOMBOS focuses on using expression values as its main criterion for building modules, resulting in clear expression changes in tight coexpression. It can easily extract prominent coexpression behaviour in the expression data, but might miss modules with less significant coexpression patterns. The coexpression patterns retrieved by COLOMBOS can be broadly interpreted as genes being functionally related as their expression is altered in similar ways in response to various stimuli. This functional relationship could imply coregulation (which might be identified using motif detection algorithms, see also Figure 4.1), but it is not a necessary prerequisite. Instead, it cannot be excluded that several regulatory programs might be responsible for the observed coexpression patterns. On the other hand, DISTILLER, as a global method, tries to recover distinctive modules that can be directly linked to a shared regulatory program, i.e. of which the genes are coregulated by the same (set of) regulator(s). Combining motif information with expression data, it successfully retrieves less prominent coexpression patterns with biological significance by utilizing extra information regarding the presence (or prediction, depending on the nature of the motif input data) of transcription factor binding sites in its genes promoter regions. The case study of *sodA* presented in this chapter illustrates this complementarity of these two approaches for retrieving biologically relevant results.

In this chapter, we focus on the specific application of COLOMBOS and DISTILLER. However, versatile tools as these have many other functionalities and application domains. In COLOMBOS, various options exist to build the module based on the information other than a given set of genes. Furthermore, if desirable, the module data can be downloaded for further analysis. As an example, the data of the module 'sodA_c3_g.75' is exported into file 'expdata_COLOMBOS_module_information.txt' (available as a sample file, see Section 4.2.6). DISTILLER can use other data sources as input, for instance, the information on the binding of regulators as obtained from ChIP-chip experiments [80, 81]. Furthermore, it can be used together with data source other than motif

data to discover other type of networks, as long as connections exist between the data source and the gene expression profile. Two such example data sets are protein-protein interaction data or synthetic lethality data.

Chapter 5

MAGIC: access portal to a cross-platform gene expression compendium for maize

5.1 Introduction

Owing to the importance of maize as sustainable food and feedstock, maize genomics is of high academic and industrial relevance. As a result, microarrays have been widely applied to interrogate the maize transcriptome, with currently over one hundred maize gene expression experiments being publicly available in online repositories such as GEO [9] and ArrayExpress [101]. However, cross-platform differences, the lack of consistent platform and measurement descriptions, and inconsistent gene annotations complicated the straightforward use of these data.

To integrate the data from different array platforms in a readily usable single compendium, we resolved gene annotation inconsistencies by reannotating probes of previously published *Zea mays* arrays using the published maize genome sequence [112] and made measurements comparable across different

This work has been published in **Q. Fu**, A.C. Fierro, P. Meysman, A. Sanchez-Rodriguez, K. Van depoele, K. Marchal, K. Engelen, MAGIC: access portal to a cross-platform gene expression compendium for Maize, *Bioinformatics* 2014

platforms/experiments using an adapted version of the data integration method described in Chapter 2. This resulted in a cross-platform expression compendium containing 1749 microarrays covering 24690 genes. Additional gene information was integrated from various external sources. Experimental annotations were manually curated. A web access portal MAGIC with a specialized set of exploration and analysis functionalities provides public access to this compendium.

5.2 Materials and Methods

5.2.1 Compendium creation

The compendium itself corresponds to a matrix of which the rows correspond to genes and columns to sample contrasts. A sample contrast is defined as the comparison of the gene‘s expression between two different biological samples, one acting as a test and the other as a reference. Each value in the matrix then represents the expression change of a gene presented as the log-ratio of its expression in the test versus the reference samples.

Probe reannotation

Probes were reannotated using the latest release of the maize genome. Original probe sequences, if available, were obtained from the respective platform annotation files or otherwise from GenBank based on the corresponding GI numbers. They were used as query against the curated ‘Filtered Gene Set’ (FGS) of the 5b.60 release of B73 maize genome using MegaBLAST version 2.2.17 [144]. Both the gene and the transcript models of FGS are searched to increase the eventual hit rate. Different blast parameters were chosen for oligo and cDNA probes, respectively, as they considerably differ in length and specificity. For probes that mapped to multiple genes, we identified the most specific hit by comparing the hit qualities across different targets. Only probes for which a sufficiently unique probe-to-gene mapping (details in Appendix B.1) could be identified were retained for compendium construction.

Expression data homogenization

The expression compendium is created utilizing our compendium creation methodology (in Chapter 2) integrated with the customized probe annotation

of our own. First, microarray experiments were retrieved from GEO and ArrayExpress. Raw expression intensities were extracted for each channel (sample) of an array separately. The intensities were subsequently normalized using dedicated procedures and mapped to the corresponding genes based on our own probe annotation. Proper test and reference samples from the same experiment were assigned based on their corresponding annotations to form contrasts. Finally, to make measurements between single- and dual-channel platforms comparable, all expression intensities were converted to log ratios (as contrasts), which compare the expression between the test and reference samples, to form the compendium. Special strategies have been developed to handle data generated using Affymetrix platform and multiple-chip platforms[1]. The details are explained in Appendix B.2.

5.2.2 Compendium annotation

To improve biological interpretation of the compendium, we integrated gene (row) and contrast (column) annotation from publicly available resources and curated all available information. To facilitate gene selection, we included, next to gene ids from 5b.60 genome release, gene names from MaizeGDB [78] and Xref assignments from www.maizesequence.org to provide mapping from EntrezGene and UniProt ids. As for functional annotations, metabolic pathway information (version 2.0) from Gramene [139] and Gene Ontology annotations from www.maizesequence.org were provided.

To compensate for the often cryptic and incomplete condition annotations available in public expression repositories, we provided curated annotations incorporating the information from both online repositories and the corresponding publications. Note that in our compendium, expression values are represented as log-ratios of a contrast between two samples. Experimental annotation is provided both at the level of the individual sample and that of the contrast. Annotation at the sample level includes tissue, development stage and genotype specifications (breeding line). The first two are described using Plant Ontology [3] derived ontology terms, whereas genotype specifications are based on the names of cultivars or wild-type. At the contrast level, we associated perturbation annotations specified as a set of relevant properties and corresponding values of change that reflect the stimuli that trigger expression alterations in the test versus the reference.

[1] Multiple-chip platform are multiple microarray chips that are designed together with their probes targeting complementary gene sets, and that are used in combination to interrogate the same biological sample in order to measure the expression levels of more genes than would be possible with only one chip.

5.2.3 Compendium exploration

To facilitate the exploration of the compendium, we constructed a web access portal MAGIC providing a set of analysis functionalities. At first, MAGIC allows users to specify their own subcompendium of interest, which only contains contrasts sharing the same characteristics. Users can choose between various predefined subcompendia. Subcompendia focusing on environmental perturbations and on comparisons between lines, between tissues and between development stages are available. Alternatively, users can generate customized subcompendia based on the sample and contrast annotations.

Once a (sub)compendium is selected, the system provides tools to explore and visualize the expression data in a module-centralized manner where a module is defined as a subset of the (sub)compendium containing the expression values of a set of genes under a set of contrasts. A module can be created starting from a query set of genes or contrasts to which contrasts or genes are added, respectively, based on their properties (such as the coexpression level, or the expression consistency). An existing module can easily be altered, merged with other modules or split into several modules. Each module can be visualized as an interactive heatmap accompanied by the corresponding annotation information.

Missing value handling

Most of the microarray platforms that MAGIC relies on were developed before the genome release of maize. None of them cover the full FGS gene set, and overlap in measured genes can be low for some platform combinations (Supplementary Tables B.6 and B.8). Consequently, each contrast contains some missing values, as not all genes are measured in the corresponding platforms.

Several strategies are employed to tackle this issue. While creating the compendium, FGS genes without any measurement are removed. We further extended the web port with extra functionalities that help users to evaluate and control the number of missing values in their analysis results. In every module creation and modification functions that select(or remove) genes or contrasts, the missing value information is provided for each candidate, and can be utilized to filter out those having too few measurements. However, under certain situations, it is not possible to directly control the candidates based on missing values, for example, manually adding genes into a module. Alternatively, the percentage of the missing value is calculated for each module. Users can consult this information and take further action if necessary. By integrating these functionalities into the web portal, the amount of missing values allowed in the user data can be easily contained. Detailed explanations

for those functionalities are provided in an online help section dedicated to this issue.

5.3 Results and Discussion

5.3.1 The compendium and the MAGIC web portal

The compendium available through MAGIC currently contains 24690 genes and 1310 sample contrasts. It covers 62% of the genes in FGS of the 5b.60 release of B73 maize genome; the remaining genes are not represented on any of the 27 platforms included. The contrasts consist of 69 experiments obtained from GEO and ArrayExpress, amounting to 1749 microarrays. Details on the composition of the compendium in terms of the number of genes and experiments covered per platform can be found in Tables B.6 and B.7. On average, a gene is measured in 9 of the 26 platforms and has been measured in 592 of 1310 contrasts (Figures B.2 and B.3).

In addition to a large volume of expression data, the compendium is also complemented with a broad range of other types of information. The latter includes extensive gene functional annotations. In total, 22812 Gene Ontology annotations are incorporated, containing 609 distinct GO terms over 15203 genes, including 11521 genes (46.7%) for which we currently have measurements in the compendium. Information on a total of 467 metabolic pathways is included into the compendium (version 2.0 from Gramene), and 3982 genes are annotated as belonging to at least one pathway. Among them, 3113 genes (12.6%) covering 429 pathways have measurements in the present compendium. Additionally, 143084 Xref references are integrated into the compendium so that users can search genes of interests by EntrezGene ids or UniProt ids. On top of this, great efforts are taken to manually curate extensive formal annotations for sample contrasts in the compendium. The samples included in the compendium belong to 30 distinct tissues obtained at 37 developmental stages from the plants of 104 different breeding lines. A total of 74 different perturbations were applied on about half of the sample contrasts currently contained in the compendium (593 out of 1310). Four predefined subcompendia are created based on these sample contrast annotations, containing respectively 488, 177, 207, 593 contrasts in the line comparison subcompendium, the tissue comparison subcompendium, the development stage comparison subcompendium, and the perturbation subcompendium. These rich sets of annotations are further incorporated into MAGIC web portal to facilitate targeted data exploration and interpretation.

The whole compendium can be downloaded through its web portal MAGIC. Alternatively, users can also directly explore the compendium utilizing various data exploration and visualization functionalities available in MAGIC. An elaborate online help, together with two tutorial case studies, illustrates how the various functionalities of MAGIC can be used to infer new biology.

5.3.2 Case studies

Two case studies are provided in MAGIC to demonstrate how interesting biological discoveries could be made using the functionalities available in the system.

Case study 1 One of the important functionality of a compendium is to provide clues for the function of unknown genes through the guilt-by-association inference based on coexpression. In the first case study we exploit this application of our compendium. To this end, 11 seed genes of unknown function were taken from those in Module 9 (hereinafter referred as M9) reported on the original work of Ficklin et al. (2011) [45]. In this study, they constructed a maize gene coexpression network from public expression data (Affymetric platform data only) in order to discover the molecular subsystem underlying complex traits. In a subsequent step, they partitioned the resulting coexpression network into several coexpression modules, one of which is module M9. This module comprises six genes of unknown functions and a set of 36 annotated genes enriched on histone and DNA binding functions. All genes in M9 are highly coexpressed. This makes M9 appealing for testing the guilt-by-association inference strategy based on other genes that coexpress with them.

In this case study, the six unknown genes of M9 were accompanied with five additional unknown genes that are directly connected to (and coexpressed with) M9 genes in the coexpression network depicted by Ficklin et al. (2011) but that do not belong to M9. This was done to increase the number of query genes and therefore reduce the influence of missing values to attain a better MAGIC performance. Starting from these 11 seed genes, we first extracted 144 contrasts form the full compendium under which these seed genes show the most prominent and consistent expression changes, and then identified over 500 genes sharing similar expression profiles with the seeds. A module named 'M9-no.anno-Mcut-gRM.gExt' was created from these contrasts and genes.

The data of the 144 contrasts of the 'M9-no.anno-Mcut-gRM.gExt' module were generated on seven different platforms with 65 out of 144 contrasts being from the Affymetrix platform. GO enrichment analysis of genes belong to this module identifies three enriched GOslim plant terms: GO:0006259 (DNA metabolic process), GO:0016043 (cellular component organization), and GO:0034645 (cellular macromolecular synthetic process). As GOslim terms specify rather general biological functions, we further analyzed the contribution of GO annotations of individual genes to those enriched terms, and discovered that genes annotated by the following

four GO terms contributes most for the observed enrichment. They are GO:0006260 (DNA replication), GO:0006270 (DNA replication initiation), GO:0006334 (nucleosome assembly), and GO:0006352 (transcription initiation). These GO terms are closely related to the histone and DNA binding functions of M9 identified by Ficklin et al (2011).

In the second part of this case study, we created the module 'M9-genes' using the other 36 known genes of M9 as seeds. The goal is two folds: to check how well 'M9-no.anno-Mcut-gRM.gExt', which is initialized with the unknown genes related to M9, captures those 36 known genes of M9 based on our compendium data, and to directly compare the similarity between module 'M9-no.anno-Mcut-gRM.gExt' and 'M9-genes'. Using module overlap visualization function to show them together, we made several observations. First, 11 out of 36 known genes are retrieved by module 'M9-no.anno-Mcut-gRM.gExt'. Second, both modules share 88 contrasts, which is 61% and 66.7% of all contrasts of 'M9-no.anno-Mcut-gRM.gExt' and 'M9-genes' respectively. At last, the missing values (unmeasured genes by some platforms) might be the main reason why two thirds of known genes of M9 (25 out of 36) are missed out by 'M9-no.anno-Mcut-gRM.gExt', although their known expression profiles are very similar to that of the other 11 genes retained in the module. Additionally, the two modules also share 4 out of 5 enriched GO terms. Our analysis has shown that coexpressed genes identified in M9 by Ficklin et al. (2011) are also coexpressed under even a broader range of conditions that are not included in the original study. And the guilt-by-association inference based on coexpression in our compendium has hinted on biological functions related to those enriched in M9. The result obtained provided extra evidence that supports their findings.

Case study 2 This case study demonstrated how MAGIC could be used as an exploratory tool to identify interesting gene candidates for further research. The goal of the case study is to identify leaf specific genes that are differentially expressed in leaves when compared with other tissues.

To do so, we manually selected leaf tissue related contrasts in the tissue comparison subcompendium and let the system identify genes that show prominent expression changes under those contrasts. A coexpression module 'gC1.3' was obtained, which comprises 82 contrasts and 66 genes. The identified genes are highly related to photosynthesis, which is evident from the enrichment on photosynthesis GO terms and two relevant pathways: C4 photosynthetic carbon assimilation cycle, oxygenic photosynthesis, despite the limited number of genes involved[2]. Furthermore, a significant number of genes (7) are described to be Chlorophyll related. As the leaf is the organ where the photosynthesis process takes place in the plant, it is to be expected that photosynthesis related genes are over-expressed in the leaves when compared with other tissues. However, after extending the module with more contrasts where genes are differentially expressed ('gC1.3_cE'), a close examination of their expression profiles revealed that these genes could be further subdivided into two sets

[2]Maize genes are poorly characterized. Among 66 genes in the module, only 15 have pathway annotations, 38 have GO annotations, and 19 are hypothetical protein.

based on their expression behaviors under a set of abiotic stress related contrasts: the over-expressed and the under-expressed genes. The observed discrepancy in the expression behavior between these two sets of genes might indicate that they are controlled by different regulatory mechanisms, hence have different biological roles, although both over-expressed when normal leaves are compared with other tissues.

Additional analysis (not included in the tutorial) had confirmed that genes in the under-expressed gene set are enriched for the photosynthesis function, whereas those over-expressed ones are related to the stress and/or stimulus response, which well explains their over-expression in the presence of various abiotic stresses. The same general stress genes are also over-expressed in normal leaf tissues. As reviewed by Baier and Dietz (2005) [5], photosynthesis process imposes a high level of oxidation stress to the leaf tissue. Consequently, the stress related genes are actively recruited to reduce this thread, resulting in their high expression level in leaf tissues.

These two case studies clearly show that the compendium expression data are of high quality, and the functionalities provided in MAGIC web portal are very useful to retrieve biologically relevant information from such a compendium. Additionally, they also illustrate how information from different microarray platforms contributes to the results obtained, and how to cope with missing values that could become abundant when the information from certain sets of platforms is combined. Step-by-step instructions are provided in detail, so users can follow them to repeat the proposed analysis. Additionally, the complete data set of each case study can be directly loaded into the workspace (from the tutorial page).

5.3.3 Discussion

In contrast to MAGIC, comparable initiatives treat data from different platforms or experiments separately. Genevestigator [58] and CORNET [31] construct separate compendia for the Affymetrix Maize Genome Array (GPL4032) (containing 558 and 340 arrays for Genevestigator and CORNET) and the Nimblegen Maize 385 k Array (GPL12620) (containing 180 arrays in both systems). PLEXdb [30], on the other hand, provides access to the data from 44 Affymetrix and Nimblegen experiments. In this system, data derived from each experiment are treated separately instead of being merged in a larger compendium. Of all these systems, only CORNET provides a restricted meta-analysis tool that allows combining information across the different compendia. Compared with these related initiatives, our approach is unique in directly combining data from different platforms in a single compendium, obviating the need for an additional meta-analysis step [46] and enabling the construction of a much larger compendium and the direct data analysis across different platforms and experiments.

Chapter 6

Conclusions and Perspectives

6.1 Summary and achievements

Chapter 2 illustrated a novel methodology to create an organism-specific cross-platform expression compendium. The uniqueness of the methodology lies in the capability of integrating expression data across different platforms. Special attention has been paid on two aspects, resolving data representation heterogeneity, particularly those related to the data generated on dual-channel microarrays, and improving data consistency and compatibility. The method has two advantages over the single platform approach. First, it facilitates the creation of a compendium that incorporates far more data than existing one, providing a more comprehensive gene expression landscape for a species. Second, it enables the construction of a sizable compendium for species without dominant microarray platform. Moreover, a web system named COMMAND has been developed, providing user friendly interfaces and guidance to facilitate compendium creation and maintenance. The system is the only known one that is capable of partially automating the tedious microarray expression data retrieval task, enabling it to handle a large volume of data. The utility of the methodology has been proven by successfully creating several bacteria compendia [38, 93], and one eukaryotic compendium for monocot *Zea mays* [49].

Chapter 3 presents three comprehensive organism-specific cross-platform expression compendia for the bacterial model organisms (*Escherichia coli*, *Bacillus subtilis*, and *Salmonella enterica* serovar Typhimurium), and the accompany web access portal COLOMBOS. Each compendium also incorporates

extensive annotations for genes as well as experimental conditions; these heterogeneous data are integrated in the COLOMBOS analysis tools to interactively browse and query the compendia. The web portal has been directly utilized in various studies, to identify additional gene coexpressed with targets [48, 91, 94], to identify conditions under which genes of interests are coexpressed [33]. The compendia, which can be downloaded in entirety and studied utilizing different systems biology approaches [32, 80, 95, 142, 146], have been incorporated in diverse researches, to construct co-expression networks [25, 74], to reconstruct transcriptional regulatory networks [43], to understand the physiological mechanism driving the response to environmental changes [7], and to study expression conservation and divergence between species [92].

Chapter 4 discusses how to discover condition dependent co-expression modules containing specific query genes in expression compendia utilizing two complementary methods COLOMBOS and DISTILLER. The former is designed for query-driven interactive data explorations in expression compendia alone, whereas the latter generates a global regulatory network overview through integrating expression data with extra evidence, e.g. motif data, in an unsupervised fashion. Both methods generate biologically relevant yet distinctive modules for the query gene *sodA*. The case study demonstrates that COLOMBOS is most optimal to extract prominent coexpression behavior among functionally related genes, whereas DISTILLER, guided with the motif information, recovers co-regulated genes albeit with a less prominent coexpression patterns. Their applications hence are driven by the type of the biological question asked. The case study will surely alleviate the difficulty faced by biologists to choose between both methods.

Chapter 5 describes an expression compendium for *Zea mays* integrating large amount of publicly available data (1749 microarrays in 69 experiments over 27 platforms). Uniquely, the probe sequences of all 27 platforms included are obtained, and a complete probe re-annotation based on the *Zea mays* 5b.60 genome release is constructed. Incorporating this re-annotation to build the compendium greatly improves the consistency between the data obtained of from platforms of different origin. Additionally, an extended condition annotation system reflecting the complex lifestyle of plant are developed, specifying not only the external perturbations at the contrast level, but also the internal sample attributes, including genotype (breeding line), tissue, and developmental stage. This compendium is made available through an upgraded web portal MAGIC that hosts a variety of analysis tools utilizing the extended annotations for easy data browsing and analysis. The uniqueness and the high quality of the maize compendium coupled with the friendly system to explore it will surely make this a valuable resource.

In summary, the main contributions of the research work presented in this thesis

are twofold. First, we developed a unique data integration methodology to create cross-platform expression compendium from publicly available data, and developed a system to facilitate the creation of such an compendium. Second, we have created such compendia for several species and developed web portals to serve them to the community.

The compendium and the system have been employed in various studies to directly generate or provide support for new biological discoveries. The *E. coli* and *B. subtilis* compendia have been employed by respectively Lemmens *et al.* (2009) [80] and Abeer *et al.* (2009) [40] to study the condition dependence and modularity of bacterial transcriptional network. Several novel targets of the regulator Fnr predicted in [80] were experimentally validated. In a research studying mutation rate plasticity (MRP) in *E. coli*, Krasovec *et al.* (2014) used our compendium to study the expression correlations between genes of interests under a broad range of experimental conditions in order to reveal the underlying mechanism behind the observed MRP [75]. Furthermore, the bacterial compendia have also been employed to study the expression conservation and divergence between species to understand how organism-specific environments drive the divergence of expression among genes conserved between closely related species [92], and to shed light on the conservation and divergence of the underlying regulatory networks across species [142, 143]. Expression data covering diverse experimental conditions that are readily available in our compendia have greatly simplified and encouraged such large scale studies. On the other hand, the targeted exploration of compendium data has also provided valuable biological insights in various studies. De Smet *et al.* (2011) have employed the *E. coli* compendium to analyze the ChIP-chip data of an independent study to help distinguish non-functional from functional bindings [32]. In this study, the obtained query-driven biclusters provide evidence for one third of the targets identified by the ChIP-chip experiment. Similarly, the same compendium data has also provided evidence to support novel transcription-factor binding sites predictions based on structural DNA properties [94]. In an evolutionary study on *Salmonella*, the *S.* Typhimurium compendium and COLOMBOS web portal have been employed to study genes gained and lost by the most recent common ancestor of *S. enterica subsp. enterica*, providing clues for the possible functions of those genes that, otherwise, have little or no homology to known functions [33]. The diversity of these applications has proven that our expression compendia are valuable resources to answer a broad range of research questions.

6.2 Future perspectives

Possible extensions of current compendium system

Compendium creation and curation The first and foremost task is to keep the existing compendia up-to-date, by updating them with newly generated datasets and expanding their annotations. New revisions of the existing compendia will be updated every half a year incorporating additional experiments. When the genome annotation of the corresponding species is revised, a new release will be generated. Upon the public release of COLOMBOS v2.0 [93], the existing compendia for the bacterial model organisms *E. coli*, *B. subtilis*, and *S. enterica* serovar Typhimurium have been extensively expanded, and four new compendia for the bacteria *Streptomyces coelicolor*, *Pseudomonas aeruginosa* *Mycobacterium tuberculosis*, and *Helicobacter pylori* have been added.

Based on novel next generation sequencing (NGS) technology, RNA sequencing (RNA-seq) quantifies relative genes expression level by directly sequencing the expressed transcripts then estimating the abundance from the generated reads. In contrast to the hybridization based microarray, such a sequence-based method does not rely on pre-existing knowledge of gene sequences, instead is capable of determining the actual sequence of RNA and quantifying expression level for individual isoforms of genes. Moreover, it has been shown that RNA-seq data have higher sensitivity, less variation, and a broader dynamic range than data obtained with microarrays [88, 117]. Due to its superiority and dropping application cost, RNA-seq is quickly replacing microarrays to become the standard method for gene expression studies. It is then crucial that our system should support this type of data. A straightforward approach has been implemented in COLOMBOS v2.0, in which the RNA-seq data are first mapped to genome, then genes relative expression levels are estimated from the mapped reads, and at last, log-ratios are calculated based on predefined sample pairs. The validity of the approach has been shown using real experimental data. The approach has enabled a swift incorporation of RNA-seq data into the existing compendia. Further research is needed to revise the existing database model to include other details that are omitted in the current approach, such as, the assembled sequence of each RNA molecular, the expression level of individual gene isoforms, etc. Novel data exploration, analysis, and visualization methods need to be developed to handle expression information of both gene and its isoforms under a common framework. Moreover, both NGS and RNA-seq are still the emerging and rapidly changing fields. We will continuously evaluate new computational methods upon available and incorporate them when appropriate to improve the data quality.

The automation of various tasks for raw data collection often involves straightforward keywords matching and pattern recognition, the rich free text descriptive information is ignored by the current system. The lack of a training set has prevented us to use text mining technologies to analyze this information. Now, successfully parsed data provide an ample training set to make this feasible. The text mining can then be employed to develop a system that assists manual annotation curation by suggesting relevant properties through analyzing meta data retrieved from the online repositories. A similar system ZOOMA [147] has been developed by EBI for such purpose. However, it maps only onto Experimental Factor Ontology (EFO) term used by EBI databases.

Compendium exploration and visualization As exemplified in this research work, a cross-platform compendium provides comprehensive expression profiles not only for different genes but also for a variety of environmental and genetic variations (contrast). One interesting feature would be allowing users to review their specific study from the respective of existing knowledge. A straightforward approach could be allowing user to upload their own data and compare it to what is available in the compendium to identify contrasts possessing similar expression variations. Such an analysis has been exemplified in the early compendium paper [60]. Alternatively, relevant experiments selected based on annotations could be visualized together with their own work to investigate diverse mechanisms underlying a common phenotype.

Initially, the web access portal only allows data to be visualized as (possibly overlapping) heatmap. The function to visualize genes and the corresponding functional annotations as an interactive network for individual module has been added in COLOMBOS v2.0 [93]. Data visualization for systems biology is a booming field, in which a variety of tools with different focus exists [52]. It will be interesting to explore this rich resource to either incorporate advanced methods into our system or link out to external services to provide better visualization and improve the data interpretability.

Expression beyond mRNA

Except for the messenger RNA which ultimately translated into protein, there are many other RNA molecules that are transcribed in a cell. They are commonly referred as non-coding RNA (ncRNA), as they do not code for proteins. Among them, transfer RNAs and ribosome RNAs have been well studied, as they carry out crucial biological functions in a cell and are well conserved across the tree of life. In the last decades, the functionality of other ncRNAs has gradually been revealed. Long non-coding RNAs (lncRNAs) are found to take up many regulatory roles, being transcriptional, post-transcriptional, or even

epigenetic[6]. Micro RNA (miRNA) induced silencing has been identified as one of the major post-transcriptional regulation mechanism that is well preserved among eukaryotes[20]. Due to the important roles played by ncRNAs in the transcriptional and post-transcriptional regulations, it is desirable to adapt the compendium creation methodology to incorporate their expression information into the compendium, providing a more comprehensive picture of transcription. As a proof of concept, by identifying probes targeting sRNA through a platform reannotation, a special *E. coli* compendium containing only data generated on the *E. coli* Antisense Genome Array and the *E. coli* Genome 2.0 Array has been created to study sRNA-mRNA interactions and to extend known regulatory network with post-transcriptional networks [66].

Comparative transcriptomics

Comparative genomics is a field in which the genomic features of different organism are compared to study the evolutionary mechanisms and the phylogeny among organisms. Such study has enabled knowledge transfer between species and facilitated gene annotation and regulatory elements identification. Traditionally, such study has focused primarily on analyzing sequence-based features, including gene sequences, gene order (synteny, genetic linkage), regulatory sequences, protein sequences, protein domains, etc. Integrating functional genomics information, such as, expression data, in the comparative study has been shown to provide new insights for study conservation and divergence [11]. The availability of comprehensive compendia for multiple bacterial species in COLOMBOS has facilitated research that studies expression conservation and divergence between *E. coli* and *S. enterica* sev. Typhimurium at global level [92]. Alternatively, algorithms, such as COMODO [142] and cMonkey [131], do exist that can simultaneously explore heterogeneous expression data sets of multiple species to identify biclusters containing conserved core orthologous genes. The mehodology presented in this thesis will enable the creation of such comprehensive compendia for many species, which can benefit the comparative genomics study of various scopes. Moreover, it will be very interesting to develop web systems that facilitate such analysis using existing algorithms, and novel methods to analyze and visualize the results obtained.

Compendium beyond transcriptome

A cell is a complex and dynamic system of which the phenotypes are driven by the interplay between molecules at different layers, including, DNA, RNA, protein, metabolites, etc. Global profiles of molecules of a single-layer and/or certain character of them only provide incomplete observations for such a system.

Combining knowledge of different aspects in research, however, provides a comprehensive and integrated view of cellular machinery, so that more precise system level models can be built to greatly reduce the false positive rates of the resulting predictions. Such an integrative approach has been successfully utilized to develop novel cis-regulatory module prediction algorithms of improved precision [56, 122], to better detect cancer-specific molecular alternations [133], and to generate integrative personal omics profile (iPOP) providing a wealth of information for personalized disease diagnosis and treatment [21]. A comprehensive cross-platform expression compendium, created using the method developed here, will remain as a premium data source to support this kind of research. Furthermore, it is of great interests to extend the methodology to construct multi-dimensional compendia incorporating different types of information, and to develop both novel data exploration methods to directly analyze such a high-dimensional data set and visualization methods to intuitively and interactively present the obtained results.

Appendix A

Appendix A: expression compendium exploration functionalities

COLOMBOS provides rich functionalities to create and/or edit expression 'modules'. In chapter 4, we explored the option that let COLOMBOS automatically identify relevant contrasts based on the user specified query gene(s). Alternatively one can also specify the conditions in interests and letting COLOMBOS automatically identify the genes that are over-expressed or under-expressed in them. Furthermore, it is also possible to maintain the full control by specifying both the genes and the conditions in interests in order to check the behavior logged in the existing experimental data.

When specifying genes of interests, apart from manually inputting gene information, one can also select genes based on other annotations obtained from public databases, such as the transcription factor or sigma factor that regulates the gene's transcription [50], the pathway a gene belongs to [71], or the transcription unit a gene belongs to [71]. These functionalities are available as options in the gene selection section of the module creation panel.

Additionally, several alternative functions are provided to choose the contrasts besides the condition selection based on expression values as used here. In case one is interested in specific experiments, one can select them directly. When a user is interested in specific types of condition properties, a condition hierarchy is available. The properties are grouped into 4 major categories: Genomic, Growth, Medium, and Treatment. Subcategories exist under each

major category to further classify them. This functionality can be reached through option 'By annotation' in the condition selection section of the module creation panel. One can also use the condition property ontology in a similar way by selecting 'By ontology' in the condition selection section.

For automatic expression value based contrasts or genes selection procedures, the calculations used to score the relevance of a contrast for a set of genes, the similarity of genes across a set of contrasts, the variability of a gene across a set of contrasts, and the Gene Ontology term enrichment are explained in this section.

Even more options are available when the user modifies an existing module. To check all available functionalities, we refer the online help section available at COLOMBOS site.

A.1 Contrast relevance score

The default relevance score c of a condition contrast for a group of genes is calculated as the absolute inverse coefficient of variation of those genes' expression values in this contrast. It is defined as the absolute mean divided by the standard deviation of the genes' expression values:

$$c = \frac{|\mu|}{\sigma} \tag{A.1}$$

On the one hand, for expression values of the same mean, the higher the score, the less sparse the values are. It prioritizes the contrasts where genes' expression values are more consistent. On the other hand, for expression values of the same standard deviation, the higher the score, the higher the mean. It prefers the contrasts where genes are highly expressed. The score thus serves as a measure that values both magnitude of expression change in response to a condition contrast, as well as coherence of expression within that contrast. The score identifies the most relevant contrasts as those where the genes 'act as one', showing the same, preferentially large, magnitude of expression change with individual variations ideally only constituting random Gaussian noise. From this notion, the score represents the number of standard deviations the mean expression value of this distribution is situated away from 0, and can be interpreted as a Z-score for the selected genes' expression change as a whole. (Note that in case of only one gene, the score of a condition contrast is degraded to the absolute expression value of that gene under it.)

We have also provided an alternative measure for the contrast relevance (selectable by the box in the top right of the contrast selection window) called 'M value cutoff'. The score assigned to the contrasts in this case is the minimum M value (i.e. the log-ratio) for the considered genes in case all genes' M values are positive, or the maximum absolute M value in case all genes' M values are negative. In case of both positive and negative M values exist for the considered genes, the contrast gets a score of 0.

A.2 Gene similarity score

The default similarity between a gene and a module's mean profile is the *Uncentered Pearson's correlation* calculated based on the formula:

$$r_v = \frac{1}{n}\sum_{i-1}^{n}(\frac{x_i}{\sigma_x^{(0)}})(\frac{y_i}{\sigma_y^{(0)}}) \tag{A.2}$$

where $\sigma_x^{(0)} = \sqrt{\frac{1}{n}\sum_i x_i^2}$; and $\sigma_y^{(0)} = \sqrt{\frac{1}{n}\sum_i y_i^2}$ Here x_i represents the candidate gene expression data at condition contrast i, whereas y_i represents a module's mean expression value at contrast i. $\sigma_x^{(0)}$ and $\sigma_y^{(0)}$ are both uncentered standard deviations assuming zero mean of the population, hence they are marked with superscript '(0)'. The higher the v_r, the more similar the expression profile of a gene is to a module's mean expression profile.

We have also provided an alternative measure for the gene similarity (selectable by the box in the top right of the contrast selection window) that calculates an uncentered version of the *Spearman rank correlation*. The Spearman rank correlation is calculated in the same way as the Pearson correlation but on the ranks of the data instead of the data itself. To calculate an uncentered version, instead of ranking all values from low to high, the positive log-ratios are ranked from low to high while the negative log-ratios are ranked from high to low and then assigned a negative sign; the mean rank is assumed 0. This uncentered Spearman rank correlation, compared to the uncentered Pearson correlation, has the advantage of being able to capture non-linear similarities, but the disadvantage of ignoring the actual magnitudes of expression changes and their distributions.

For ranking genes, there are three options provided based on the uncentered (rank) correlation score calculated. First one is 'positive' which uses v_r directly as final score. The second option 'absolute' takes $|v_r|$. It ranks both correlated

and anti-correlated genes based solely on their similarities. Instead, the third option 'negative' takes $-v_r$ as a score to favor only the anti-correlated genes.

A.3 Gene variability

The variability of a gene's expression value x for conditions $i = 1, \ldots, n$ is calculated as the uncentered standard deviation:

$$\sigma_x^{(0)} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2} \tag{A.3}$$

A.4 Enrichment calculation

The enrichment score p-value is calculated based on a *hypergeometric distribution*. Given a genome of size N, there are K genes in it associated with a Gene Ontology (GO) term T, the p-value representing the chance to observe k or more such genes appear in a random module of size n is calculated as follows:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{A.4}$$

The lower the p-value, the more significant a GO term is enriched in the module.

Appendix B

Appendix B: Magic supplementary methods

B.1 Preprocessing: probe to gene mapping

A semi-automatic workflow has been developed to consistently annotate probes (Figure B.1), i.e. to identify a unique target gene for each probe whenever possible. In total we needed to map 209036 probes, originating from 27 different microarray platforms. Target genes belong to the "Filtered Gene Set" (FGS) of 5b RefGen v2 B73 maize genome release, since it contains only the high quality gene predictions by removing possible pseudogenes, transposons, contaminations, and low confidence genes. Both the FGS Gene Model and the FGS Transcript Model are used in our workflow, in order to achieve the highest possible mapping coverage for assigning probes to their proper target genes. The Gene Model contains full gene sequences, including exons, introns, 3' UTRs, etc, while the Transcript Model contains only transcript sequences, including splice variants. The workflow consists of four major steps, as is illustrated in Figure B.1. First, the collected probe sequences are BLASTed against both the gene model and transcript model. Next, one-to-one probe mappings are extracted by taking all unique hits and identifying the top- hits from multiple hits. The corresponding quality scores are calculated. In the third step, results from the Gene Model BLAST and Transcript Model BLAST are merged into a consistent probe to gene map by resolving possible conflicts between one probe's gene hit and transcript hit based on the comparison of their quality scores. At last, the mappings retained in previous step are subjected to an additional filtering step

to remove low quality hits. Note that we only do quality filtering in the final step in order to maximize the information retained to identify and resolve the potential ambiguous probe sequences. The results are a high quality one-to-one probe to gene mapping.

The results of the workflow are influenced by the characteristics of the input probe sequences, which serve as BLAST query sequences in step 1. We make a distinction between oligo and cDNA probes (respectively 158694 and 60345 in total),. Oligo probe sequences are short sequences of length less than 100 nucleotides, usually sifted through a stringent selection process [82, 108]. In contrast, cDNA sequences (which we retrieved from NCBI GenBank based on the access id referred by each probe in the platform specifications), are much longer sequences with length varying between one hundred to several thousand bases. Often generated as a single-pass read, they are of varying quality, and some contain low complexity regions in their sequence. The differences between these two groups are reflected by the parameters used when applying our workflow on them. In the initial BLAST step, an e-value cutoff 0.001 is used for oligo due to their shorter length. In contrast, a much stricter e-value cutoff 1e-20 is applied for cDNA to avoid hits over low quality regions and to compensate their longer sequence length. Conversely, a stricter criterion for oligos is employed to guarantee the mapping quality in the final filtering step, as even small variances between probe and target sequences can have a great influence on their binding specificity due to the short sequence length. A looser criterion is utilized for cDNA assuming that longer probe sequences can tolerate more sequence variation and still bind the proper target transcripts.

In the next sections, the individual steps of the workflow, and the results obtained from each step, will be discussed in greater detail.

B.1.1 Step 1 – Mapping with megablast

First, the probe sequences (BLAST queries) are blasted against both the gene model and transcript model (BLAST targets) using megablast version 2.2.17 ([144]). BLAST on both the FGS Gene Model and Transcript Model was done to recover as much of the tentative targets of each probe, because the collected probe sequences, especially the cDNA ones, sometimes contain also introns. In addition, for certain sequences, BLAST on the transcript model alone will result in poor quality hits, and as a consequence, for many probes no target gene can be identified. To retain as much information as possible from the blast results we choose a relative loose criterion to BLAST sequences. Except for the different e-value cutoffs, the common parameters applied for both cDNA

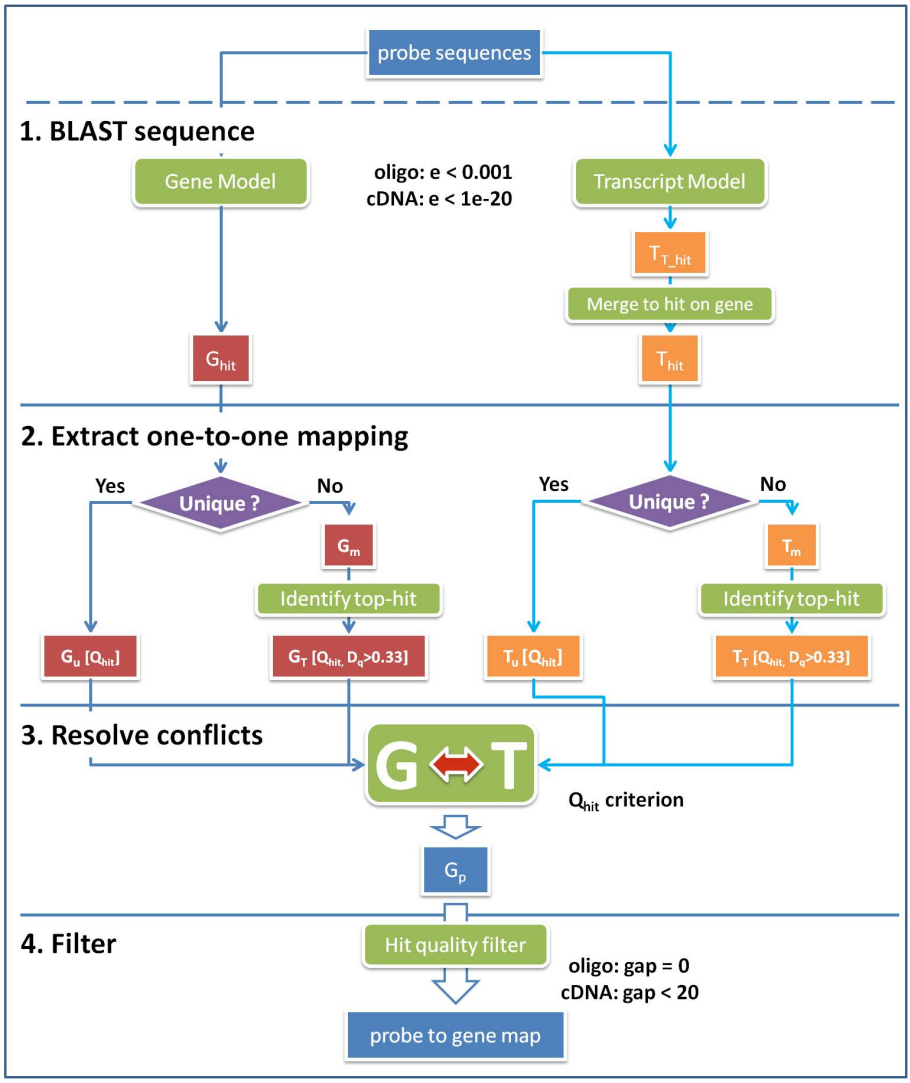


Figure B.1: Probe to gene mapping workflow. The workflow consists of four steps. First, the probe sequences collected are BLASTed against both the FGS Gene and Transcript Model. Next, one-to-one probe mappings are extracted by taking all unique hits (G_u, T_u) and identifying top-hits (G_T,T_T) from multiple hits (G_m,T_m). For all hits the quality measurements Q_{hit} and D_q are calculated. In the third step, results from Gene Model BLAST and Transcript Model BLAST are merged into a consistent probe to gene map (G_p) by resolving possible conflicts between one probe's gene hit and transcript hit using Q_{hit}. At last, G_p is filtered to remove low quality hits, resulting in a high quality one-to-one probe to gene mapping.

Table B.1: Probe mapping for cDNA and oligo probes for Zea mays

	cDNA		Oligo	
	Gene blast	**Transcript blast**	**Gene blast**	**Transcript blast**
Probe total[2]		60345		158694
Step 1 - Mapping with megablast				
Hits total[3]	153050	130927	129097	123089
Hit transcripts[4]	-	42250	-	49289
Hit genes[5]	24495	23393	28879	28416
Probes retained	57700	56852	99139	105429
Step 2 - Extracting one-to-one mappings				
Unique hits[6]	24766	24820	84854	92878
Multiple hits				
Hits count[3]	128284	106107	44243	30211
Probe count	32934	32032	14285	12551
Top hits[7]	9067 (27.5%)	12979 (40.5%)	11 (0.0%)	11 (0.0%)
Step 3 - Merging blast results[8]				
Unique maps[9]	898	40	7697	13987
Identical maps[10]		52084		90482
Conflicts[11]		2647 (4718)		6 (953)
Sum[12]		38563 (*57740*)		97001 (*113119*)
Step 4 - Filtering by quality				
Probes removed		414		6870
Result				
Contribution[13]	4305 (11.3%)	33844	5187 (5.8%)	84944
Total		38149		90131

1 If not noted, the number of unique probes of a corresponding category is reported in the table.
2 The same set of probe sequence is used in both the gene blast and the transcript blast procedure.
3 The number of BLAST hits of a corresponding category is reported.
4 T_{T_hit}, the number of unique transcripts having at least one hit is reported.
5 The number of unique genes having at least one hit is reported.
6 For unique hits where a probe hits only on one target (gene/transcript), these three numbers are equal, the number of hits, the number of unique targets, and the number of the unique probes.
7 The number of probes and the corresponding percentages that the top hits pass D_q criterion.
8 In this step, the hits from different blasts are merged to form one-to-one probe to gene mappings. To contrast the results with those obtained in previous steps, each record in them are called a *map* instead of a *hit* in the later steps.
9 Probes whose targets are only identified by one blast procedure. Further breakdowns of the data based on hit type are available in Table B.2.
10 Probes for which both gene blast and transcript blast identify the same targets. Further breakdowns of the data into different categories are available in Table B.3.
11 Probes for which different targets are identified by gene blast and transcript blast (a conflict), showing the number of resolved conflicts with the total number of conflicts between brackets. Please check Section B.1.3 for details, and a data breakdown in Table B.3.
12 The total number of probes left in the final merged mappings. In bracket, the raw total count before removing ambiguous top hits.
13 The number of probes obtaining maps from each blast analysis.

sequences and oligo sequences are '$-F\,F$' to turn off the query sequence filtering, '$-b\,15$' to list only the top 15 hits.

BLAST on the Gene Model generates a set of hits G_{hit}, each of which maps a probe to a gene; BLAST on the Transcript Model generates a different set of hits T_{T_hit}, which maps a probe to a transcript. T_{T_hit} are converted into hits on genes in two steps. First, all hits of a probe to a transcript are grouped by the corresponding genes of the hit transcripts. Second, each group of hits is merged into one hit on that gene, while the best transcript hit score in the group is retained as the gene hit score. After this conversion, both Gene Model BLAST results G_{hit} and the Transcript Model BLAST results T_{hit} map probes to gene identifiers. Nevertheless, to distinguish between the gene and transcript hit, we keep referring to the probe-to-gene hits originating from the Transcript Model BLAST (T_{hit}) as the probe-to-transcript hits.

After applying this step, the results obtained are summarized in the Table B.1 (Step 1). The number of the transcript hits is nearly the double of the number of the gene hits. When mapped to a gene, many probes in each category hit on the different transcripts of a gene (data not shown), indicating that they are not designed to distinguish different transcripts (splice variants) of the same gene.

B.1.2 Step 2 – Extracting one-to-one mappings

In this step, we try to extract one-to-one probe to gene mapping from both the G_{hit} and T_{hit} lists independently. The gene hit and transcript hit results are kept separately to better resolve the possible conflicts between them in the next step. Based on the number of identified target genes a probe sequence has, both G_{hit} and T_{hit} are divided into unique-hit group and multi-hit group, where a single probe maps to only one gene or on several genes respectively. This results in 4 groups:

- G_u, gene unique-hit
- T_u, transcript unique-hit
- G_m, gene multi-hit
- T_m, transcript multi-hit

Next, we calculate a hit quality score Q_{hit} for each hit, based on its BLAST information as follows:

$$Q_{hit} = coverage * identity - 3 * num\_gaps/query\_length \tag{B.1}$$

in it, *coverage*, *identity*, and *num_gaps* (number of gaps) are characteristics of the BLAST hit, and 3 is empirically chosen such that enough penalty is given to gaps but not overweight it so much that Q_{hit} can become negative. As a simple percentage, the score takes into the consideration the percentage of exact match nucleotides on probe sequence and the number of gaps in the match region. Those are important factors that influence probe to target specificity. We use this scores to resolve the multiple mapping issues in the multi-hit groups (G_m and T_m) by identifying a promising best hit for each probe -if possible-. First, all hits of a probe are ranked by their Q_{hit}s. The hit with the highest score are kept, resulting in G_t, the gene top-hits, and G_t, the transcript top-hits. Then, the difference D_q between the scores of the top two hits is calculated. It serves as a proxy for the binding specificity difference between first two hits. When the following condition is met:

$$D_q = Q_{hit_{1st}} - Q_{hit_{2nd}} \geq 0.33 \tag{B.2}$$

the hit with the highest score (top hit) is assumed to be the target of the probe, i.e. considered more likely to bind the probe sequence compared to other hits. If the above condition is not met, the corresponding results for that probe are marked, assuming that they can hybridize several genes and generate ambiguous expression measurements. They are kept temporarily to identify possible conflicts between gene and transcript blast output, and to resolve the conflicts by comparing hits quality.

The result of this step is summarized in Table B.1 (Step 2). Clearly, the short oligo probe sequences are much more target specific when compared with the cDNA sequences, with the majority being unique hits and much less multiple mapping probes. In contrast, cDNA probe sequences, although much longer, tend to produce partial hits on several genes, due to the relative loose BLAST settings in the step 1. Filtered by the D_q criterion, the true target gene (top hit) can be identified in many cases (27.5% of G_m and 40.5% of T_m). Whereas for oligo probes, this is rather rare (11 cases in both G_m and T_m).

B.1.3 Step 3 – Merging blast results

Till now we have identified one target gene for each probe separately for gene blast analysis and transcript blast analysis. Next we need to combine these two sets into one consistent result set. There are different cases when combining two data sets. The first case is the '*unique map*', where the result is obtained in only one blast analysis. In this case, the blast data are taken directly into the final data set. When both blast analyses identify the same gene as target,

Table B.2: Statistics of uniquely mapped probes

Probe blast hit type	cDNA	oligo
Gene blast		
G_u	344	5609
G_t	554	2088
Sub total	898	7697
Transcript blast		
T_u	40	11937
T_t	10	2050
Sub total	50	13987

Table B.3: Breakdown of the queries that have hits in both gene and transcript blast

		cDNA		oligo	
Gene hit	**Trans. hit**	**Identical**	**Conflict**	**Identical**	**Conflict**
G_u	T_u	23542	0	79152	7(0)
G_t	T_t	26788	4354 (579, 1885)	10102	313(0)
G_u	T_t	821	59 (4, 17)	72	14(0)
G_t	T_u	933	305 (147, 15)	1156	626(6, 0)

it is the case of '*identical map*'. The genes identified are taken as the target, and the best blast hit raw data are retained for further analysis. At last there is '*conflict*' when two blast analysis identify different genes as the target of a probe. In this case, an extra step is taken to resolve the conflict by identifying the possibly most reliable target out of the two candidates when possible. If this fails, the probe is discarded from the result set. The detail of the conflict resolving strategy is explained in the next section.

The statistic data of this analytical step is summarized in Table B.1 (Step 3), showing the probe count for each aforementioned cases. For the conflicts, it shows the probe count for the resolved cases with the total number of conflicts between brackets. Recall that in previous step, we only marked the ambiguous top-hits. Those ambiguous results retained in the merged result sets are removed before the next step. This results a big reduction of the cDNA probes mapped, dropping from 57740 to 38563 (33.2% less). Whereas for oligo probes, there is a 14.2% drop, removing 16118 out of 113119 probes. (Table B.1).

Resolving conflicts

There is a *conflict* if for one probe, the target gene of the transcript blast hit (T_u/T_t) are different from that of the gene blast hit (G_u/G_t). In previous steps, we grouped the results obtained by each blast analysis into two sets, unique hit and top hit. These results in 4 groups in total, G_u, T_u, G_t, and T_t. The conflicts among them are resolved according to a set of heuristic rules. Note that by definition, there are no conflicts between the result sets obtained from the same blast results (gene or transcript model), such as (G_u, G_t) and (T_u, T_t). For each conflict, the following condition based on Q_{hit} is evaluated:

$$ABS(Q_{T_{hit}} - Q_{G_{hit}}) \geq 0.2 \tag{B.3}$$

when true, the one with the higher Q_{hit} was chosen to be the real target gene; otherwise, the hits of corresponding probe are discarded from the results due to having ambiguous target genes. Note, the Q_{hit} cutoff used here is less strict than the one used in identifying top-hit from multiple mapping probes, since here we compare two potential hits from different BLAST results, while before hits of same BLAST results were compared.

Depending on the sources between which the conflict arises, there are three types of conflicts:

- Conflicts between a pair of unique-hits, i.e. from (G_u, T_u). There are no such conflicts for the cDNA hits, and 7 conflicts for the oligo hits. Checking their gene unique hit results, we found that all those hits reside fully or partially in the intron region. Consequently, those genes could not be identified as targets by blast against the transcript model. Similarly, the unique transcript hits are across exon boundaries of hit genes, and as such they do not appear in the gene model blast results. After applying our Q_{hit} criterion, none passed the check and all were discarded.
- Conflicts between one unique-hit and one top-hit, i.e. between (G_u, T_t) or (G_t, T_u). There are many such conflicts for both the cDNA probes and the oligo probes. By applying the above Q_{hit} condition, nearly half of the conflicts can be resolved for the cDNA probes. However, for the oligo probes, this succeeds in only 6 cases (Table B.3). And the resolved cases are mostly won by the top hits. In the table B.4, one example is given for each subtype where the conflict is resolved.
- Conflict between a pair of top-hits from (G_t, T_t). For the cDNA probes, many such conflicts exist (Table B.3). When applying the Q_{hit} condition on them, we resolved 56.6% of them, in which transcript hits win most.

A small number of such conflicts exist for the oligo probes. However, none can be successfully resolved. Such an example is shown in the table B.5. Gene 'GRMZM5G854499' has a Q_{hit} of 0.967118 as the top hit of gene blast (first row). This score is much higher than that of the best transcript hit (0.31947 on gene 'GRMZM2G162184'). Hence the former is identified as the real target of the probe. Note that both gene blast and transcript blast identified the same two genes as top two hits, although in different order. Indeed, for this type of conflicts, often the same two genes are competing for the best target of a probe.

After merging gene blast output with transcript blast output, the result set contains only one-to-one probe-to-gene mappings with high specificity to guarantee a reliable biological interpretation of their measurements.

B.1.4 Step 4 – Filtering by hit quality

As mentioned in Step 1, a loose criterion is used for BLAST in order to retain as much information in the further steps of this workflow. As a result, some hits in the merged set could still be of low quality. In this step, an filter is applied on each individual probe checking the quality of the hit based on the

Table B.4: The conflicts between a top hit and an unique hit

	Hit type	**Target**	**Coverage**	**Match length**	**Gaps**	***e*-value**
Case 1	2^{nd}	GRMZM2G020553	61.8267	250	4	$1.00E-107$
	G_t	**GRMZM5G865576**	94.61358	403	2	0
	T_u	GRMZM2G020553	61.8267	250	4	$1.00E-108$
Case 2	G_u	AC206201.3_FG004	14.61412	89	0	5.00E-43
	2^{nd}	AC206201.3_FGT004	26.76519	162	3	5.00E-79
	T_t	**GRMZM2G003109**	73.23481	389	7	1.00E-111

Table B.5: The top hits conflict example

		Target	Q_{hit}	**Coverage**	**Match Length**	**Gaps**	***e*-value**
Gm	1^{st}	**GRMZM5G854499**	**0.967118**	100	509	3	0
	2^{nd}	GRMZM2G162184	0.31947	48.743	211	15	6e-36
Tm	2^{nd}	GRMZM5G854499	0.119923	12.766	65	1	1e-26
	1^{st}	GRMZM2G162184	0.31947	48.743	211	15	4e-36

corresponding blast information. Due to sequence differences, different cutoffs were applied on oligo and cDNA probes. For short oligo probe sequences, a gap or a mismatch can have a great influence on the binding specificities of the target sequences. Hence, the filter ($num\_gaps = 0$ and $identity >= 95$) is applied. This removes 6870 probes. For the much longer cDNA sequences, a looser filter ($num\_gaps =< 20$ and $identity >= 80$) is used removing 414 probes (Table B.1).

Summary

After applying this workflow, we successfully identified the target genes for 56.8% of oligo probes and 63.2% of cDNA ones. Without compromising the mapping quality between probe and target gene sequences, our blast analysis against the Gene Model made the significant contribution to the final results, providing 4305 mappings (11.3%) for cDNA probes and 5187 mappings (5.8%) for oligo ones.

Although ideally each probe should produce a hit on only one target gene, the fact that 36% of cDNA results come from the top-hit identified from multiple gene mappings shows that the reality is far from ideal, and it is very important for a probe mapping flow to handle multiple mapping issue. Whereas only 13 out of 90131 oligo probes produce hits in top-hit lists, which demonstrates evidently that the strict probe sequence selection processes ensure good probe specificity, and result in a more reliable biological interpretation of their measurements.

B.2 Expression data retrieve and normalization

In the process of collecting data to create maize compendium, we encounter two issues which require special strategies to handle them. In the following sections, these issues and the corresponding solutions are explained.

B.2.1 Affymetrix data retrieve

For the Affymetrix Maize Genome Array with GEO access number GPL3042 and ArrayExpress A-AFFY-77, the gene expression measurements are often reported in GEO and ArrayExpress as the values summarized at the probeset level. As explained in section 2.2.2 of chapter 2, for Affymetrix data, the raw probe intensities are preferred then this summarized values. Additionally, in order to store the probe annotation of an Affymetrix microarray required to

handle raw probe intensities, the concept of 'Virtual Platform' is introduced. The 'Virtual platform' for the Affymetrix Maize Genome Array is the 'maize' platform. It stores the probe level annotation of this microarray extracted from the corresponding CDF file downloaded from Affymetrix website.

B.2.2 Multiple-chip platform data normalization

Zea mays is a complex Monocots with a very large genome. Because the older array design did not have enough capacity to cover the full gene set using a single microarray, multiple chips of the same technology, each with their own probes targeting complementary gene sets were used. These are referred as the multiple-chip platform. The complementarity of this platform is utilized by hybridizing the same biological sample on multiple chips of it to obtain expression data for a extended set of genes. Consequently, the data generated on this type of platform requires special handling in the annotation and the homogenization step to generate the normalized data for compendium. The details are explained in the corresponding parts in section 2.2.2 of chapter 2.

B.3 Supplementary Tables and Figures

Table B.6: Platform data overview

PlatformID	Data source	Type	Probe count	Gene count[1]	Exp. count	Contr. count	Name
GPL372	GEO	cDNA	8895	1639	2	37	ZmDB 606-Immature Ear Microarray (2 cm)
GPL498	GEO	cDNA	10182	3124	3	50	ZmDB Array Unigene–1–01–01
GPL499	GEO	cDNA	10362	3124	2	20	ZmDB Array Unigene-1-01-07
GPL1208	GEO	cDNA	10362	3124	1	8	Zea mays Unigene01_01_04
GPL1990	GEO	cDNA	15053	11750	2	14	Maize oligo array version 1.2 array A
GPL1991	GEO	cDNA	15220	12157	2	14	Maize oligo array version 1.2 array B
GPL1992	GEO	cDNA	15053	11750	4	68	Maize oligo array version 1.3 array A
GPL1993	GEO	cDNA	15220	12157	4	68	Maize oligo array version 1.3 array B
GPL2557	GEO	cDNA	11569	6909	2	36	SAM1.0
GPL2572	GEO	cDNA	8757	6207	3	42	SAM2.0
GPL2613	GEO	cDNA	11559	6910	1	54	SAM1.1
GPL2984	GEO	cDNA	9488	4069	1	18	Maize Unigene 1-02-01
GPL3021	GEO	cDNA	8222	6298	1	144	ISU Maize 12k cDNA Generation II Version B-IG
GPL3099	GEO	cDNA	12841	9447	1	24	Agilent Maize 21K v1.0
GPL3333	GEO	cDNA	11569	6909	3	34	SAM1.1a
GPL3538	GEO	cDNA	9991	7023	4	72	SAM3.0
GPL3618	GEO	Affy	2122	1386	1	23	Maize CornChip0 8.5K GeneChip
GPL4521	GEO	cDNA	9561	5970	1	9	Maize SAM1.2 Array
GPL5439	GEO	cDNA	15053	11750	10	160	Maize oligo array version 1.9 array A
GPL5440	GEO	cDNA	16008	12595	10	160	Maize oligo array version 1.9 array B
GPL6053	GEO	cDNA	8121	6407	1	36	Maize 12K cDNA Generation II Version B.1
GPL6092	GEO	cDNA	10362	3124	1	8	MGDP Zea mays Unigene 01_01_05
GPL6438	GEO	cDNA	24856	17540	4	118	Maize oligonucleotide array 46K version
GPL6460	GEO	cDNA	21832	15948	1	12	Universidad Nacional de Rosario Zea Mays 43K

Continued on next page

Table B.6 – Platform data overview (continued)

PlatformID	Data source	Type	Probe count	Gene count[1]	Exp. count	Contr. count	Name
GPL7209	GEO	Agilent	25184	16987	1	14	Zea mays 1x44K Agilent array - designed by Walbot Lab
GPL7297	GEO	cDNA	12673	9470	1	36	Zea mays 22K Agilent array, Ver2 - designed by Walbot Lab
Maize	Internal[2]	Affy	396398	10291	25	345	Affymetrix Maize Genome Array [Maize]

1 This count is the number of unique gene ids probed by a platform. A gene measured by multiple probes is counted as only once.

2 Platform 'Maize' refers to the probe level annotation of the Affymetrix Maize genome array (GEO platform GPL3042 and ArrayExpress platform A-AFFY-77). The 'Maize' label is internally used by MAGIC to differentiate the probe-level chip information from the probe set level information that is already provided in GPL3042 and A-AFFY-77.

Table B.7: Experiment data overview

Experiment Id	Data source	Contrast count	Sample count	Platforms	Multi-chip platform[1]
GSE573	GEO	27	54	GPL372	
GSE671	GEO	22	44	GPL498, GPL499	
GSE1353	GEO	8	16	GPL1208	
GSE1807	GEO	36	72	GPL498, GPL499	
GSE2163	GEO	12	24	GPL498	
GSE2771	GEO	10	20	GPL372	
GSE3017	GEO	144	288	GPL3021	
GSE3490	GEO	18	36	GPL2984	
GSE3640	GEO	24	48	GPL3099	
GSE3890	GEO	12	24	GPL1990, GPL1991, GPL1992, GPL1993	*
GSE4466	GEO	10	20	GPL3333	
GSE4477	GEO	54	108	GPL2613	
GSE4663	GEO	23	24	GPL3618	
GSE6267	GEO	18	36	GPL2557,GPL2572, GPL3538	*
GSE7030	GEO	3	4	Maize[2]	
GSE7248	GEO	30	60	GPL2572, GPL3333, GPL3538	
GSE8188	GEO	16	18	Maize[2]	
GSE8194	GEO	33	33	Maize[2]	

Continued on next page

Table B.7 – Experiment data overview (continued)

Experiment Id	Data source	Contrast count	Sample count	Platforms	Multi-chip platform[1]
GSE8308	GEO	23	24	Maize[2]	
GSE8320	GEO	47	48	Maize[2]	
GSE9379	GEO	24	48	GPL1990, GPL1991, GPL1992, GPL1993	*
GSE9386	GEO	12	24	GPL5439, GPL5440	*
GSE9430	GEO	36	72	GPL6053	
GSE9453	GEO	30	32	GPL1992, GPL1993	*
GSE9546	GEO	8	16	GPL6092	
GSE9610	GEO	18	36	GPL2557, GPL2572, GPL3538	*
GSE9698	GEO	12	24	GPL5439, GPL5440	*
GSE10023	GEO	35	36	Maize[2]	
GSE10236	GEO	26	27	Maize[2]	
GSE10237	GEO	9	9	Maize[2]	
GSE10243	GEO	7	8	Maize[2]	
GSE10308	GEO	16	32	GPL1992, GPL1993	*
GSE10400	GEO	12	24	GPL6460	
GSE10449	GEO	4	8	GPL6438	
GSE10542	GEO	12	24	GPL6438	
GSE10543	GEO	23	48	GPL6438	
GSE10544	GEO	54	108	GPL5439, GPL5440	*
GSE10596	GEO	4	8	GPL5439, GPL5440	*
GSE11325	GEO	36	72	GPL3333, GPL3538	
GSE11531	GEO	3	4	Maize[2]	
GSE12579	GEO	14	28	GPL7209	
GSE12756	GEO	36	72	GPL7297	
GSE12892	GEO	6	6	Maize[2]	
GSE13768	GEO	9	18	GPL4521	
GSE15048	GEO	4	4	Maize[2]	
GSE15371	GEO	5	6	Maize[2]	
GSE16567	GEO	22	24	Maize[2]	
GSE17754	GEO	63	126	GPL6438	
GSE17932	GEO	8	16	GPL5439, GPL5440	*
GSE17953	GEO	16	32	GPL5439, GPL5440	*
GSE17971	GEO	11	22	GPL5439, GPL5440	*
GSE18006	GEO	13	26	GPL5439, GPL5440	*
GSE18008	GEO	12	24	GPL5439, GPL5440	*
GSE18011	GEO	18	36	GPL5439, GPL5440	*
GSE18491	GEO	8	9	Maize[2]	

Continued on next page

Table B.7 – Experiment data overview (continued)

Experiment Id	Data source	Contrast count	Sample count	Platforms	Multi-chip platform[1]
GSE19501	GEO	6	8	Maize[2]	
GSE19559	GEO	3	3	Maize[2]	
GSE19883	GEO	16	32	GPL6438	
GSE21070	GEO	22	24	Maize[2]	
GSE22479	GEO	10	12	Maize[2]	
GSE24624	GEO	9	10	Maize[2]	
E–MEXP–1222	ArrayExpress	11	12	Maize[2]	
E–MEXP–1464	ArrayExpress	5	6	Maize[2]	
E–MEXP–1465	ArrayExpress	5	6	Maize[2]	
E–MEXP–2364	ArrayExpress	5	6	Maize[2]	
E–MEXP–2366	ArrayExpress	5	6	Maize[2]	
E–MEXP–2367	ArrayExpress	5	6	Maize[2]	
E–MEXP–2368	ArrayExpress	5	6	Maize[2]	
E–MEXP–2702	ArrayExpress	7	8	Maize[2]	
Total		1310	2255		
	GEO	1262	2199		
	ArrayExpress	48	56		

1 Multiple-chip platform are multiple microarray chips that are designed together with their probes targeting complementary gene sets, and that are used in combination to interrogate the same biological sample in order to measure the expression levels of more genes than would be possible with only one chip. Data from the same biological samples but generated on multiple chips of this platform are combined in our system.

2 Platform 'Maize' refers to the probe level annotation of the Affymetrix Maize genome array (GEO platform GPL3042 and ArrayExpress platform A-AFFY-77). The 'Maize' label is internally used by MAGIC to differentiate the probe-level chip information from the probe set level information that is already provided in GPL3042 and A-AFFY-77.

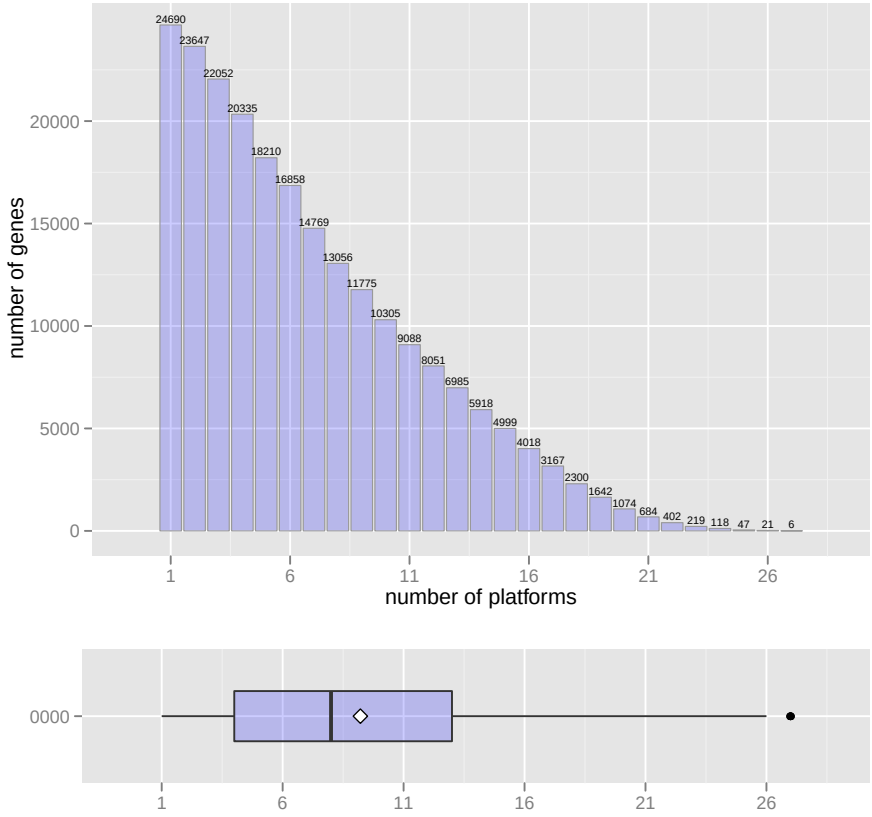
Table B.8: Overlap in gene content between pairs of platforms

	GPL1208	GPL1990 GPL1991[1]	GPL1992 GPL1993[1]	GPL2557 GPL2572 GPL3538[1]	GPL2572	GPL2613	GPL2984	GPL3021	GPL3099	GPL3333	GPL3538	GPL3618	GPL372	GPL4521	GPL498	GPL499	GPL5439 GPL5440[1]	GPL6053	GPL6092	GPL6438	GPL6460	GPL7209	GPL7297	Maize
GPL1208	**3124**	2777	2777	2530	1307	1904	407	1704	1970	1920	1339	342	299	1510	3124	3124	2792	1740	3124	2734	2311	2531	1968	1916
GPL1990 GPL1991	2777	**19967**	**19967**	**10492**	5456	6088	3632	5597	8435	6091	6076	1249	1463	5257	2777	2777	**19967**	5693	2777	**17327**	**13349**	**14214**	8442	8808
GPL1992 GPL1993	2777	**19967**	**19967**	**10492**	5456	6088	3632	5597	8435	6091	6076	1249	1463	5257	2777	2777	**19967**	5693	2777	**17327**	**13349**	**14214**	8442	8808
GPL2557 GPL2572 GPL3538	2530	**10492**	**10492**	**12246**	6207	6858	3573	5648	7122	6909	7023	1161	1422	5780	2530	2530	**10575**	5636	2530	**10164**	**9133**	**9785**	7089	7043
GPL2572	1307	5456	5456	6207	**6207**	2762	1576	2782	3960	2754	3172	576	738	2731	1307	1307	5493	2789	1307	5373	4652	5015	3939	3804
GPL2613	1904	6088	6088	6858	2762	**6910**	3146	4599	4428	6795	3914	954	1167	5110	1904	1904	6136	4531	1904	5983	5209	5613	4423	4547
GPL2984	407	3632	3632	3573	1576	3146	**4069**	2963	2733	3118	2090	793	1201	2348	407	407	3665	3197	407	3579	3146	3377	2723	2919
GPL3021	1704	5597	5597	5648	2782	4599	2963	**6298**	3854	4602	3472	974	1229	3748	1704	1704	5634	6021	1704	5502	4786	5148	3848	4154
GPL3099	1970	8435	8435	7122	3960	4428	2733	3854	**9447**	4422	4048	880	1043	3732	1970	1970	8475	3968	1970	8202	7324	7907	9210	5576
GPL3333	1920	6091	6091	6909	2754	6795	3118	4602	4422	**6909**	3910	950	1177	5153	1920	1920	6140	4519	1920	5988	5209	5613	4416	4554
GPL3538	1339	6076	6076	7023	3172	3914	2090	3472	4048	3910	**7023**	792	935	3411	1339	1339	6118	3501	1339	5857	5419	5785	4042	4269
GPL3618	342	1249	1249	1161	576	954	793	974	880	950	792	**1386**	510	760	342	342	1255	1007	342	1219	1083	1156	873	1158
GPL372	299	1463	1463	1422	738	1167	1201	1229	1043	1177	935	510	**1639**	1000	299	299	1475	1263	299	1449	1263	1359	1055	1394
GPL4521	1510	5257	5257	5780	2731	5110	2348	3748	3732	5153	3411	760	1000	**5970**	1510	1510	5298	3603	1510	5160	4503	4834	3732	3887
GPL498	3124	2777	2777	2530	1307	1904	407	1704	1970	1920	1339	342	299	1510	**3124**	3124	2792	1740	3124	2734	2311	2531	1968	1916
GPL499	3124	2777	2777	2530	1307	1904	407	1704	1970	1920	1339	342	299	1510	3124	**3124**	2792	1740	3124	2734	2311	2531	1968	1916
GPL5439 GPL5440	2792	**19967**	**19967**	**10575**	5493	6136	3665	5634	8475	6140	6118	1255	1475	5298	2792	2792	**20180**	5732	2792	**17540**	**13457**	**14329**	8484	8883
GPL6053	1740	5693	5693	5636	2789	4531	3197	6021	3968	4519	3501	1007	1263	3603	1740	1740	5732	**6407**	1740	5598	4893	5256	3957	4208
GPL6092	3124	2777	2777	2530	1307	1904	407	1704	1970	1920	1339	342	299	1510	3124	3124	2792	1740	**3124**	2734	2311	2531	1968	1916
GPL6438	2734	**17327**	**17327**	**10164**	5373	5983	3579	5502	8202	5988	5857	1219	1449	5160	2734	2734	**17540**	5598	2734	**17540**	**12324**	**13134**	8212	8615
GPL6460	2311	**13349**	**13349**	**9133**	4652	5209	3146	4786	7324	5209	5419	1083	1263	4503	2311	2311	**13457**	4893	2311	**12324**	**15948**	**15948**	7323	7550
GPL7209	2531	**14214**	**14214**	**9785**	5015	5613	3377	5148	7907	5613	5785	1156	1359	4834	2531	2531	**14329**	5256	2531	**13134**	**15948**	**16987**	7906	8087
GPL7297	1968	8442	8442	7089	3939	4423	2723	3848	9210	4416	4042	873	1055	3732	1968	1968	8484	3957	1968	8212	7323	7906	**9470**	5668
Maize	1916	8808	8808	7043	3804	4547	2919	4154	5576	4554	4269	1158	1394	3887	1916	1916	8883	4208	1916	8615	7550	8087	5668	**10291**

1 Multiple-chip platform are multiple microarray chips that are designed together with their probes targeting complementary gene sets, and that are used in combination to interrogate the same biological sample in order to measure the expression levels of more genes than would be possible with only one chip. Data from the same biological samples but generated on multiple chips of this platform are combined in our system. The gene count for a multiple-chip platform is the number of unique gene ids probed by all chips of the platform. A gene measured by multiple chips is counted as only once.

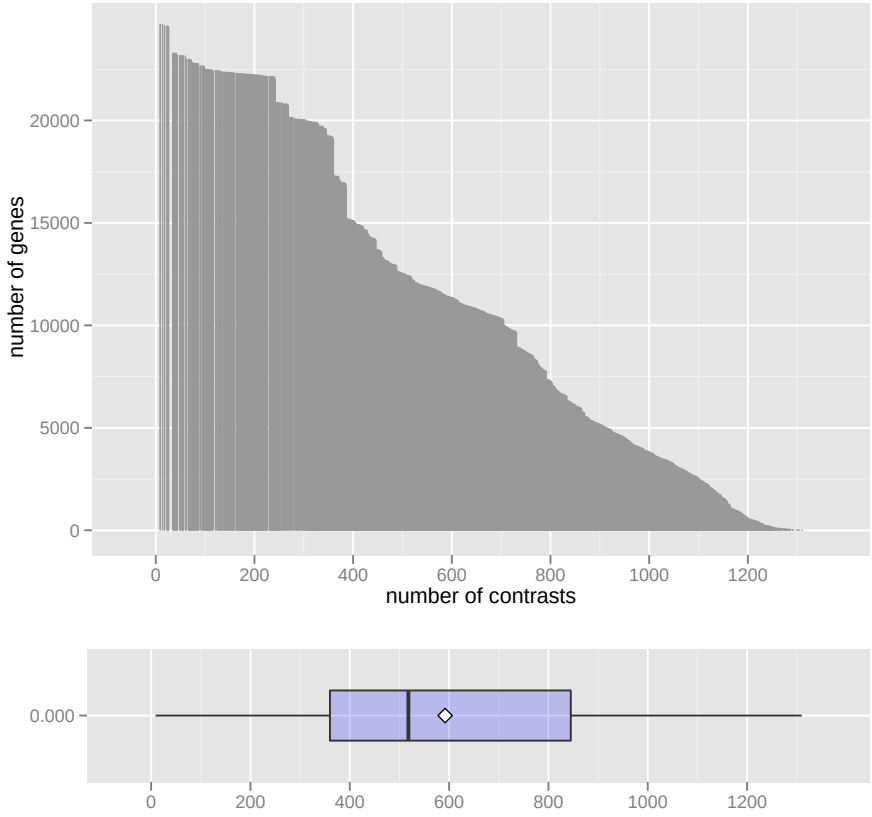
2 The shade of a cell corresponds to the number of genes shared by a pair of platform. The higher the number the darker the shade.

Figure B.2: Gene platform coverage



The figures show gene platform coverage. Above, a bar chart showing the number of genes (y-axis) covered by at least x platforms indicated by the x-axis. The leftmost bar shows that each gene is measured on at least 1 platform, whereas there are only 6 genes that have been measured on all 27 platforms of the compendium (the rightmost bar). The box plot below indicates the number of platforms a gene has been measured on (gene centric and non-cumulative counting), showing that most genes are measured on 4 to 13 platforms with an average of 9 (the diamond).

Figure B.3: Gene contrast coverage



The figures show gene contrast coverage. Above, a bar chart showing the number of genes (y-axis) covered by at least x contrasts indicated by the x-axis. The leftmost bar shows that every gene has measurements in at least a small number of contrasts (the corresponding x value of this bar is close to but not at 0). Few genes have measurements in all 1310 contrasts in the compendium (the rightmost bar). The box plot below indicates the number of contrasts a gene has been measured in (gene centric and non-cumulative counting), showing that most genes have measurements in between 300 to 900 contrasts with an average of 592 (the diamond).

Bibliography

[1] S. F. Altschul et al. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.” In: *Nucleic Acids Research* 25.17 (1997), pp. 3389–3402.

[2] L. H. Augenlicht and D. Kobrin. “Cloning and screening of sequences expressed in a mouse colon tumor.” In: *Cancer research* 42.3 (Mar. 1982), pp. 1088–93.

[3] S. Avraham et al. “The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations”. In: *Nucleic Acids Research* 36.Database issue (Jan. 2008), pp. D449–D454.

[4] M. M. Babu, B. Lang, and L. Aravind. “Methods to Reconstruct and Compare Transcriptional Regulatory Networks”. In: *Methods in molecular biology (Clifton, N.J.)* 143 (2009), pp. 1–17.

[5] M. Baier and K.-J. Dietz. “Chloroplasts as source and target of cellular redox regulation: a discussion on chloroplast redox signals in the context of plant physiology.” In: *Journal of experimental botany* 56.416 (June 2005), pp. 1449–62.

[6] M. Baker. “Long noncoding RNAs: the search for function”. In: *Nature Methods* 8.5 (May 2011), pp. 379–383.

[7] Y. I. Balderas-Martínez, M. Savageau, H. Salgado, E. Pérez-Rueda, E. Morett, and J. Collado-Vides. “Transcription factors in Escherichia coli prefer the holo conformation.” In: *PloS one* 8.6 (Jan. 2013), e65723.

[8] T. Bammler et al. “Standardizing global gene expression analysis between laboratories and across platforms.” In: *Nature methods* 2.5 (May 2005), pp. 351–6.

[9] T. Barrett et al. “NCBI GEO: archive for functional genomics data sets — 10 years on”. In: *Nucleic Acids Research* 39.Database issue (2011), pp. D1005–D1010.

[10] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. "Reverse engineering of regulatory networks in human B cells." In: *Nature genetics* 37.4 (Apr. 2005), pp. 382–90.

[11] S. Bergmann, J. Ihmels, and N. Barkai. "Similarities and differences in genome-wide expression data of six organisms." In: *PLoS biology* 2.1 (Jan. 2004), E9.

[12] D. di Bernardo et al. "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks." In: *Nature biotechnology* 23.3 (Mar. 2005), pp. 377–83.

[13] Y. Blat and N. Kleckner. "Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region." In: *Cell* 98.2 (July 1999), pp. 249–59.

[14] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." In: *Bioinformatics (Oxford, England)* 19.2 (Jan. 2003), pp. 185–93.

[15] A. Brazma. "Minimum Information About a Microarray Experiment (MIAME)–successes, failures, challenges." In: *TheScientificWorldJournal* 9 (Jan. 2009), pp. 420–3.

[16] A. Brazma et al. "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." In: *Nature genetics* 29.4 (Dec. 2001), pp. 365–71.

[17] D. Buck and J. R. Guest. "Overexpression and site-directed mutagenesis of the succinyl-CoA synthetase of Escherichia coli and nucleotide sequence of a gene (g30) that is adjacent to the suc operon." In: *The Biochemical journal* 260.3 (June 1989), pp. 737–47.

[18] E. Camon et al. "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology". In: *Nucleic Acids Research* 32.Database issue (2004), pp. D262–D266.

[19] R. Caspi et al. "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases." In: *Nucleic Acids Research* 36.Database issue (2008). Ed. by A. Regev, pp. D623–31.

[20] K. Chen and N. Rajewsky. "The evolution of gene regulation by transcription factors and microRNAs." In: *Nature reviews. Genetics* 8.2 (Feb. 2007), pp. 93–103.

[21] R. Chen et al. "Personal omics profiling reveals dynamic molecular and medical phenotypes." In: *Cell* 148.6 (Mar. 2012), pp. 1293–307.

[22] Z. Chen et al. “Discovery of Fur binding site clusters in Escherichia coli by information theory models”. In: *Nucleic Acids Research* 35.20 (2007), pp. 6762–6777.

[23] S. E. Choe, M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon. “Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.” In: *Genome biology* 6.2 (Jan. 2005), R16.

[24] W. S. Cleveland. “Robust Locally Weighted Regression and Smoothing Scatterplots”. In: *Journal of the American Statistical Association* 74.368 (1979), pp. 829–836.

[25] L. Cloots and K. Marchal. “Network-based functional modeling of genomics, transcriptomics and metabolism in bacteria.” In: *Current opinion in microbiology* 14.5 (Oct. 2011), pp. 599–607.

[26] *COLOMBOS*. URL: http://bioi.biw.kuleuven.be/colombos.

[27] I. Compan and D. Touati. “Anaerobic activation of arcA transcription in Escherichia coli: roles of Fnr and ArcA.” In: *Molecular microbiology* 11.5 (1994), pp. 955–964.

[28] P. A. Cotter and R. P. Gunsalus. “Contribution of the fnr and arcA gene products in coordinate regulation of cytochrome o and d oxidase (cyoABCDE and cydAB) genes in Escherichia coli.” In: *FEMS microbiology letters* 70.1 (1992), pp. 31–36.

[29] A. C. Culhane et al. “GeneSigDB: a manually curated database and resource for analysis of gene expression signatures.” In: *Nucleic acids research* 40.Database issue (Jan. 2012), pp. D1060–6.

[30] S. Dash, J. Van Hemert, L. Hong, R. P. Wise, and J. A. Dickerson. “PLEXdb: gene expression resources for plants and plant pathogens.” In: *Nucleic acids research* 40.Database issue (Nov. 2011), gkr938–.

[31] S. De Bodt, J. Hollunder, H. Nelissen, N. Meulemeester, and D. Inzé. “CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations.” In: *The New phytologist* 195.3 (Aug. 2012), pp. 707–20.

[32] R. De Smet and K. Marchal. “An ensemble biclustering approach for querying gene expression compendia with experimental lists.” In: *Bioinformatics (Oxford, England)* 27.14 (July 2011), pp. 1948–56.

[33] P. T. Desai et al. “Evolutionary Genomics of Salmonella enterica Subspecies.” In: *mBio* 4.2 (Jan. 2013).

[34] *DISTILLER*. URL: http://bioi.biw.kuleuven.be/DISTILLER.

[35] S. Draghici, P. Khatri, A. C. Eklund, and Z. Szallasi. “Reliability and reproducibility issues in DNA microarray measurements.” In: *Trends in genetics : TIG* 22.2 (Feb. 2006), pp. 101–9.

[36] A. Elfilali, S. Lair, C. Verbeke, P. La Rosa, F. Radvanyi, and E. Barillot. “ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis.” In: *Nucleic acids research* 34.Database issue (Jan. 2006), pp. D613–6.

[37] K. Engelen, B. Naudts, B. De Moor, and K. Marchal. “A calibration method for estimating absolute expression levels from microarray data.” In: *Bioinformatics* 22.10 (2006), pp. 1251–1258.

[38] K. Engelen et al. “COLOMBOS: Access Port for Cross-Platform Bacterial Expression Compendia”. In: *PLoS ONE* 6.7 (July 2011).

[39] J. Ernst, Q. K. Beg, K. a. Kay, G. Balázsi, Z. N. Oltvai, and Z. Bar-Joseph. “A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli.” In: *PLoS computational biology* 4.3 (Mar. 2008), e1000044.

[40] A. Fadda, A. C. Fierro, K. Lemmens, P. Monsieurs, K. Engelen, and K. Marchal. “Inferring the transcriptional network of Bacillus subtilis.” In: *Molecular bioSystems* 5.12 (Dec. 2009), pp. 1840–52.

[41] J. J. Faith et al. “Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.” In: *PLoS biology* 5.1 (Jan. 2007), e8.

[42] J. J. Faith et al. “Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata.” In: *Nucleic Acids Research* 36.Database issue (Jan. 2008), pp. D866–70.

[43] J. P. Faria, R. Overbeek, F. Xia, M. Rocha, I. Rocha, and C. S. Henry. “Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models.” In: *Briefings in bioinformatics* (Feb. 2013).

[44] W. P. Fawcett, R. E. Wolf, and R. E. W. Jr. “Genetic definition of the Escherichia coli zwf "soxbox", the DNA binding site for SoxS-mediated induction of glucose 6-phosphate dehydrogenase in response to superoxide”. In: *Journal of bacteriology* 177.7 (1995), pp. 1742–1750.

[45] S. P. Ficklin and F. A. Feltus. “Gene Co-Expression Network Alignment and Conservation of Gene Modules Between Two Grass Species: Maize (Zea mays) and Rice (Oryza sativa).” In: *Plant physiology* 156.July (May 2011), pp. 1244–1256.

[46] A. C. Fierro, F. Vandenbussche, K. Engelen, Y. Van de Peer, and K. Marchal. “Meta Analysis of Gene Expression Data within and Across Species.” In: *Current genomics* 9.8 (Dec. 2008), pp. 525–34.

[47] R. Fleischmann et al. “Whole-genome random sequencing and assembly of Haemophilus influenzae Rd”. In: *Science* 269.5223 (July 1995), pp. 496–512.

[48] Q. Fu et al. “Directed module detection in a large-scale expression compendium.” In: *Methods in molecular biology (Clifton, N.J.)* 804 (Jan. 2012), pp. 131–65.

[49] Q. Fu et al. “MAGIC: access portal to a cross-platform gene expression compendium for maize.” In: *Bioinformatics (Oxford, England)* (Jan. 2014), pp. 8–9.

[50] S. Gama-Castro et al. “RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation.” In: *Nucleic acids research* 36.Database issue (Jan. 2008), pp. D120–4.

[51] T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins. “Inferring genetic networks and identifying compound mode of action via expression profiling.” In: *Science (New York, N.Y.)* 301.5629 (July 2003), pp. 102–5.

[52] N. Gehlenborg et al. “Visualization of omics data for systems biology”. In: *Nature Publishing Group* 7.3s (2010), S56–S68.

[53] T. Gene, O. Consortium, and G. Consortium. “The Gene Ontology in 2010: extensions and refinements”. In: *Nucleic Acids Research* 38.Database issue (Jan. 2010), pp. D331–D335.

[54] A. Goffeau et al. “Life with 6000 genes.” In: *Science (New York, N.Y.)* 274.5287 (Oct. 1996), pp. 546, 563–7.

[55] C. Grosse, J. Scherer, D. Koch, M. Otto, N. Taudte, and G. Grass. “A new ferrous iron-uptake transporter, EfeU (YcdN), from Escherichia coli.” In: *Molecular Microbiology* 62.1 (2006), pp. 120–131.

[56] C. Herrmann, B. Van de Sande, D. Potier, and S. Aerts. “i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules.” In: *Nucleic acids research* 40.15 (Aug. 2012), e114.

[57] L. Hertzberg, O. Zuk, G. Getz, and E. Domany. “Finding motifs in promoter regions.” In: *Journal of computational biology : a journal of computational molecular cell biology* 12.3 (Apr. 2005), pp. 314–30.

[58] T. Hruz et al. “Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes.” In: *Advances in bioinformatics* 2008 (Jan. 2008), p. 420747.

[59] E. Hubbell, W.-m. Liu, and R. Mei. “Robust estimators for expression analysis.” In: *Bioinformatics (Oxford, England)* 18.12 (Dec. 2002), pp. 1585–92.

[60] T. R. Hughes et al. “Functional discovery via a compendium of expression profiles.” In: *Cell* 102.1 (July 2000), pp. 109–26.

[61] T. Ideker, T. Galitski, and L. Hood. “A new approach to decoding life: systems biology.” In: *Annual review of genomics and human genetics* 2 (Jan. 2001), pp. 343–72.

[62] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. “Summaries of Affymetrix GeneChip probe level data.” In: *Nucleic acids research* 31.4 (Feb. 2003), e15.

[63] R. A. Irizarry, Z. Wu, and H. a. Jaffee. “Comparison of Affymetrix GeneChip expression measures.” In: *Bioinformatics (Oxford, England)* 22.7 (Apr. 2006), pp. 789–94.

[64] R. A. Irizarry et al. “Exploration, normalization, and summaries of high density oligonucleotide array probe level data.” In: *Biostatistics (Oxford, England)* 4.2 (Apr. 2003), pp. 249–64.

[65] R. A. Irizarry et al. “Multiple-laboratory comparison of microarray platforms.” In: *Nature methods* 2.5 (May 2005), pp. 345–50.

[66] I. Ishchukov, Y. Wu, S. Van Puyvelde, J. Vanderleyden, and K. Marchal. “Inferring the relation between transcriptional and posttranscriptional regulation from expression compendia.” In: *BMC microbiology* 14.1 (Jan. 2014), p. 14.

[67] F. Jacob and J. Monod. “Genetic regulatory mechanisms in the synthesis of proteins”. In: *Journal of Molecular Biology* 3.3 (June 1961), pp. 318–356.

[68] K. W. Jair et al. “Purification and regulatory properties of MarA protein , a transcriptional activator of Escherichia coli multiple antibiotic and superoxide resistance promoters . Purification and Regulatory Properties of MarA Protein , a Transcriptional Activator of Esch”. In: (1995).

[69] T. Kacmarczyk, P. Waltman, A. Bate, P. Eichenberger, and R. Bonneau. “Comparative microbial modules resource: generation and visualization of multi-species biclusters.” In: *PLoS computational biology* 7.12 (Dec. 2011), e1002228.

[70] M. Kapushesky et al. “Gene expression atlas at the European bioinformatics institute.” In: *Nucleic acids research* 38.Database issue (Jan. 2010), pp. D690–8.

[71] P. D. Karp et al. “Expansion of the BioCyc collection of pathway/genome databases to 160 genomes.” In: *Nucleic acids research* 33.19 (Jan. 2005), pp. 6083–9.

[72] M. K. Kerr and G. A. Churchill. “Experimental design for gene expression microarrays.” In: *Biostatistics (Oxford, England)* 2.2 (June 2001), pp. 183–201.

[73] I. M. Keseler et al. “EcoCyc: a comprehensive view of Escherichia coli biology.” In: *Nucleic acids research* 37.Database issue (Jan. 2009), pp. D464–70.

[74] M. Kolář, J. Meier, V. Mustonen, M. Lässig, and J. Berg. “GraphAlignment: Bayesian pairwise alignment of biological networks.” In: *BMC systems biology* 6 (Jan. 2012), p. 144.

[75] R. Krašovec et al. “Mutation rate plasticity in rifampicin resistance depends on Escherichia coli cell-cell interactions.” In: *Nature communications* 5 (Jan. 2014), p. 3742.

[76] W. P. Kuo et al. “A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies.” In: *Nature biotechnology* 24.7 (July 2006), pp. 832–40.

[77] D. A. Lashkari et al. “Yeast microarrays for genome wide parallel genetic and gene expression analysis.” In: *Proceedings of the National Academy of Sciences of the United States of America* 94.24 (Nov. 1997), pp. 13057–62.

[78] C. J. Lawrence, Q. Dong, M. L. Polacco, T. E. Seigfried, and V. Brendel. “MaizeGDB, the community database for maize genetics and genomics.” In: *Nucleic Acids Research* 32.Database issue (2004), pp. D393–D397.

[79] C. Lazar et al. “Batch effect removal methods for microarray gene expression data integration: a survey.” In: *Briefings in bioinformatics* (July 2012).

[80] K. Lemmens et al. “DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli.” In: *Genome biology* 10.3 (Jan. 2009), R27.

[81] K. Lemmens et al. “Inferring transcriptional modules from ChIP-chip, motif and microarray data.” In: *Genome biology* 7.5 (Jan. 2006), R37.

[82] G. G. Leparc et al. “Model-based probe set optimization for high-performance microarrays.” In: *Nucleic acids research* 37.3 (Feb. 2009), e18.

[83] C. Li and W. H. Wong. “Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.” In: *Proceedings of the National Academy of Sciences of the United States of America* 98.1 (Jan. 2001), pp. 31–6.

[84] M. Lukk et al. “A global map of human gene expression.” In: *Nature biotechnology* 28.4 (Apr. 2010), pp. 322–4.

[85] J. Luo et al. “A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data.” In: *The pharmacogenomics journal* 10.4 (Aug. 2010), pp. 278–91.

[86] Q. Ma, Y. Yin, M. a. Schell, H. Zhang, G. Li, and Y. Xu. “Computational analyses of transcriptomic data reveal the dynamic organization of the Escherichia coli chromosome under different conditions.” In: *Nucleic acids research* 41.11 (June 2013), pp. 5594–603.

[87] D. Marbach et al. “Wisdom of crowds for robust gene network inference.” In: *Nature methods* 9.8 (Aug. 2012), pp. 796–804.

[88] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. “RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.” In: *Genome Research* 18.9 (2008), pp. 1509–1517.

[89] E. Massé and S. Gottesman. “A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli”. In: *Proceedings of the National Academy of Sciences of the United States of America* 99.7 (2002), pp. 4620–4625.

[90] J. P. McHugh et al. “Global iron-dependent gene regulation in Escherichia coli. A new mechanism for iron homeostasis.” In: *The Journal of Biological Chemistry* 278.32 (2003), pp. 29478–86.

[91] P. Meysman, J. Collado-Vides, E. Morett, R. Viola, K. Engelen, and K. Laukens. “Structural properties of prokaryotic promoter regions correlate with functional features.” In: *PloS one* 9.2 (Jan. 2014), e88717.

[92] P. Meysman, A. Sánchez-Rodríguez, Q. Fu, K. Marchal, and K. Engelen. “Expression Divergence between Escherichia coli and Salmonella enterica serovar Typhimurium Reflects Their Lifestyles.” In: *Molecular biology and evolution* 30.6 (June 2013), pp. 1302–14.

[93] P. Meysman et al. “COLOMBOS v2.0: an ever expanding collection of bacterial expression compendia.” In: *Nucleic acids research* 42.Database issue (Jan. 2014), pp. D649–53.

[94] P. Meysman et al. “Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli”. In: *Nucleic Acids Research* 39.2 (Jan. 2011), e6.

[95] T. Michoel, R. De Smet, A. Joshi, Y. Van de Peer, and K. Marchal. “Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks.” In: *BMC systems biology* 3 (Jan. 2009), p. 49.

[96] P. Mukhopadhyay, M. Zheng, L. A. Bedzyk, R. A. LaRossa, and G. Storz. “Prominent roles of the NorR and Fur regulators in the Escherichia coli transcriptional response to reactive nitrogen species.” In: *Proceedings of the National Academy of Sciences of the United States of America* 101.3 (Jan. 2004), pp. 745–50.

[97] A. Nandal et al. “Induction of the ferritin gene (ftnA) of Escherichia coli by Fe 2 + - Fur is mediated by reversal of H-NS silencing and is RyhB independent”. In: *Molecular Microbiology* 75.3 (2010), pp. 637–57.

[98] B. Palsson. “In silico biology through "omics".” In: *Nature biotechnology* 20.7 (July 2002), pp. 649–50.

[99] F. Pan et al. “Gene Aging Nexus: a web database and data mining platform for microarray data on aging.” In: *Nucleic acids research* 35.Database issue (Jan. 2007), pp. D756–9.

[100] E. M. Panina, A. A. Mironov, and M. S. Gelfand. “Comparative analysis of FUR regulons in gamma-proteobacteria.” In: *Nucleic Acids Research* 29.24 (2001), pp. 5195–5206.

[101] H. Parkinson et al. “ArrayExpress update — from an archive of functional genomics experiments to the atlas of gene expression”. In: *Nucleic Acids Research* 37.Database issue (2009), pp. D868–D872.

[102] S. I. Patzer and K. Hantke. “Dual repression by Fe(2+)-Fur and Mn(2+)-MntR of the mntH gene, encoding an NRAMP-like Mn(2+) transporter in Escherichia coli.” In: *Journal Of Bacteriology* 183.16 (2001), pp. 4806–4813.

[103] A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. Fodor. “Light-generated oligonucleotide arrays for rapid DNA sequence analysis.” In: *Proceedings of the National Academy of Sciences of the United States of America* 91.11 (May 1994), pp. 5022–6.

[104] K. D. Pruitt, T. Tatusova, and D. R. Maglott. “NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins”. In: *Nucleic Acids Research* 35.Database issue (2007), pp. D61–D65.

[105] J. Quackenbush. “Microarray data normalization and transformation.” In: *Nature genetics* 32 Suppl.december (Dec. 2002), pp. 496–501.

[106] D. R. Rhodes et al. “Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles”. In: *Neoplasia* 9.2 (Feb. 2007), pp. 166–180.

[107] M. E. Ritchie et al. “A comparison of background correction methods for two-colour microarrays.” In: *Bioinformatics (Oxford, England)* 23.20 (Oct. 2007), pp. 2700–7.

[108] J.-M. J.-M. Rouillard, M. Zuker, and E. Gulari. "OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach." In: *Nucleic acids research* 31.12 (June 2003), pp. 3057–62.

[109] *Sample files for Chapter 4*. URL: http://bioi.biw.kuleuven.be/DISTILLER/extra_files/2011-MiMB.BMN/samplefile_page/.

[110] R. Sásik, C. H. Woelk, and J. Corbeil. "Microarray truths and consequences." In: *Journal of molecular endocrinology* 33.1 (Aug. 2004), pp. 1–9.

[111] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray". In: *Science* 270.5235 (Oct. 1995), pp. 467–470.

[112] P. S. Schnable et al. "The B73 maize genome: complexity, diversity, and dynamics." In: *Science (New York, N.Y.)* 326.5956 (Nov. 2009), pp. 1112–5.

[113] L. Shi et al. "Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential." In: *BMC bioinformatics* 6 Suppl 2 (July 2005), S12.

[114] L. Shi et al. "The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies". In: *BMC Bioinformatics* 9.Suppl 9 (Jan. 2008), S10.

[115] L. Shi et al. "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." In: *Nature biotechnology* 24.9 (Sept. 2006), pp. 1151–61.

[116] N. Sierro, Y. Makita, M. de Hoon, and K. Nakai. "DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information." In: *Nucleic acids research* 36.Database issue (Jan. 2008), pp. D93–6.

[117] A. Sîrbu, G. Kerr, M. Crane, and H. J. Ruskin. "RNA-Seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering". In: *PloS one* 7.12 (Jan. 2012), e50986.

[118] P. Stafford and M. Brun. "Three methods for optimization of cross-laboratory and cross-platform microarray expression data." In: *Nucleic acids research* 35.10 (Jan. 2007), e72.

[119] V. Storms, M. Claeys, A. Sanchez, B. De Moor, A. Verstuyf, and K. Marchal. "The effect of orthology and coregulation on detecting regulatory motifs." In: *PloS one* 5.2 (Jan. 2010), e8938.

[120] A. I. Su et al. "A gene atlas of the mouse and human protein-encoding transcriptomes." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.16 (Apr. 2004), pp. 6062–7.

[121] A. I. Su et al. “Large-scale analysis of the human and mouse transcriptomes.” In: *Proceedings of the National Academy of Sciences of the United States of America* 99.7 (Apr. 2002), pp. 4465–70.

[122] H. Sun, T. Guns, A. C. Fierro, L. Thorrez, S. Nijssen, and K. Marchal. “Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection.” In: *Nucleic acids research* 40.12 (July 2012), e90.

[123] H. Sun, K. Lemmens, T. Van den Bulcke, K. Engelen, B. De Moor, and K. Marchal. “ViTraM: visualization of transcriptional modules.” In: *Bioinformatics (Oxford, England)* 25.18 (Sept. 2009), pp. 2450–1.

[124] A. J. Sutton, K. R. Abrams, and D. R. Jones. “An illustrated guide to the methods of meta-analysis.” In: *Journal of evaluation in clinical practice* 7.2 (May 2001), pp. 135–48.

[125] P. K. Tan et al. “Evaluation of gene expression measurements from commercial microarray platforms.” In: *Nucleic acids research* 31.19 (Oct. 2003), pp. 5676–84.

[126] B. Tardat and D. Touati. “Iron and oxygen regulation of Escherichia coli MnSOD expression: competition between the global regulators Fur and ArcA for binding to DNA.” In: *Molecular microbiology* 9.1 (1993), pp. 53–63.

[127] M. Tompa et al. “Assessing computational tools for the discovery of transcription factor binding sites.” In: *Nature biotechnology* 23.1 (Jan. 2005), pp. 137–44.

[128] D. Touati. “Sensing and protecting against superoxide stress in Escherichia coli–how many ways are there to trigger soxRS response?” In: *Redox report : communications in free radical research* 5.5 (2000), pp. 287–293.

[129] P. Uetz et al. “A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.” In: *Nature* 403.6770 (Feb. 2000), pp. 623–7.

[130] *ViTraM*. URL: http://homes.esat.kuleuven.be/~kmarchal/ViTraM/Index.html.

[131] P. Waltman et al. “Multi-species integrative biclustering.” In: *Genome biology* 11.9 (Jan. 2010), R96.

[132] P. Warnat, R. Eils, and B. Brors. “Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes.” In: *BMC bioinformatics* 6 (Jan. 2005), p. 265.

[133] J. N. Weinstein et al. “The Cancer Genome Atlas Pan-Cancer analysis project.” In: *Nature genetics* 45.10 (Oct. 2013), pp. 1113–20.

[134] E. A. Welsh, S. A. Eschrich, A. E. Berglund, and D. A. Fenstermacher. "Iterative rank-order normalization of gene expression microarray data." In: *BMC bioinformatics* 14 (Jan. 2013), p. 153.

[135] R. J. Wilde and J. R. Guest. "Transcript analysis of the citrate synthase and succinate dehydrogenase genes of Escherichia coli K12." In: *Journal of general microbiology* 132.12 (Dec. 1986), pp. 3239–51.

[136] L. Xu, A. C. Tan, D. Q. Naiman, D. Geman, and R. L. Winslow. "Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data." In: *Bioinformatics (Oxford, England)* 21.20 (Oct. 2005), pp. 3905–11.

[137] L. Xu, A. C. Tan, R. L. Winslow, and D. Geman. "Merging microarray data from separate breast cancer studies provides a robust prognostic test." In: *BMC bioinformatics* 9 (Jan. 2008), p. 125.

[138] Y. H. Yang et al. "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." In: *Nucleic Acids Research* 30.4 (2002), e15.

[139] K. Youens-Clark et al. "Gramene database in 2010: updates and extensions." In: *Nucleic acids research* 39.Database issue (Nov. 2010), pp. D1085–94.

[140] S. O. Zakharkin et al. "Sources of variation in Affymetrix microarray experiments." In: *BMC bioinformatics* 6 (Jan. 2005), p. 214.

[141] M. J. Zaki and C.-J. Hsiao. "CHARM : An Efficient Algorithm for Closed Itemset Mining". In: *2nd SIAM International Conference on Data Mining* 15 (2002), pp. 457–473.

[142] P. Zarrineh, A. C. Fierro, A. Sánchez-Rodríguez, B. De Moor, K. Engelen, and K. Marchal. "COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms." In: *Nucleic acids research* 39.7 (Apr. 2011), e41.

[143] P. Zarrineh, A. Sánchez-Rodríguez, N. Hosseinkhan, Z. Narimani, K. Marchal, and A. Masoudi-Nejad. "Genome-Scale Co-Expression Network Comparison across Escherichia coli and Salmonella enterica Serovar Typhimurium Reveals Significant Conservation at the Regulon Level of Local Regulators Despite Their Dissimilar Lifestyles." In: *PloS one* 9.8 (Jan. 2014), e102871.

[144] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. "A greedy algorithm for aligning DNA sequences." In: *Journal of computational biology : a journal of computational molecular cell biology* 7.1-2 (2000), pp. 203–14.

[145] Z. Zhang, G. Gosset, R. Barabote, C. S. Gonzalez, W. A. Cuevas, and M. H. Saier. "Functional interactions between the carbon and iron utilization regulators, Crp and Fur, in Escherichia coli." In: *Journal Of Bacteriology* 187.3 (2005), pp. 980–990.

[146] H. Zhao et al. "Query-based biclustering of gene expression data using Probabilistic Relational Models." In: *BMC bioinformatics* 12 Suppl 1.Suppl 1 (Jan. 2011), S37.

[147] *ZOOMA*. URL: www.ebi.ac.uk/fgpt/zooma.

Publications

Journal papers

Published

- **Q. Fu**, A.C. Fierro, P. Meysman, A. Sanchez-Rodriguez, K. Van depoele, K. Marchal, K. Engelen, **MAGIC: access portal to a cross-platform gene expression compendium for Maize**, *Bioinformatics* 2014
- P. Meysman, P. Sonego, L. Bianco, **Q. Fu**, D. Ledezma-Tejeida, S. Gama-Castro, V. Liebens, J. Michiels, , K. Laukens, K. Marchal, J. Collado-Vides, K. Engelen, COLOMBOS v2.0: an ever expanding collection of bacterial expression compendia. *Nucleic Acids Research*, 42(Database issue), D649?53. 2013
- P. Meysman, A. Sanchez-Rodriguez, **Q. Fu**, K. Marchal, K. Engelen, Expression Divergence between Escherichia coli and Salmonella enterica serovar Typhimurium Reflects Their Lifestyles, *Molecular biology and evolution* 30, 1302?14, 2013
- K. Engelen*, **Q. Fu***, P. Meysman, A. Sanchez-Rodriguez, R. De Smet, K. Lemmens, A.C. Fierro, K. Marchal, COLOMBOS: access port for cross-platform bacterial expression compendia, *PLoS One*, volume 6, issue 7, 2011 (*These authors contributed equally to this work)
- H. Sun, T. De Bie, V. Storms, **Q.Fu**, T. Dhollander, K. Lemmens, A. Verstuyf, B. De Moor, K. Marchal, ModuleDigger: an itemset mining framework for the detection of cis-regulatory modules, *BMC Bioinformatics*, volume 10, issue 1, S30 pages, 2009

In preparation

- N. Verstraeten, C. Kint, V. Liebens, **Q. Fu**, C. Davids, A.C. Fierro, K. Marchal, J. Beirlant, J. Hofkens, M. Jansen, M. Fauvart, J. Michiels. A Conserved GTPase Controls Bacterial Persistence by Modulating the hokB-sokB Toxin-Antitoxin Module

Book chapter

- **Q. Fu**, K. Lemmens, I. Thijs, P. Meysman, A. Sanchez-Rodriguez, H. Sun, A.C. Fierro, K. Engelen, K. Marchal. Directed module detection in a large-scale expression compendium, Van Helden, J.; Toussaint, A.; Thieffry, D (eds.), *Methods in Molecular Biology - Bacterial Molecular Networks (MMB)*, 2012

FACULTY OF BIOSCIENCE ENGINEERING
DEPARTMENT OF MICROBIAL AND MOLECULAR SYSTEMS
CENTRE OF MICROBIAL AND PLANT GENETICS
Kasteelpark Arenberg 20, bus 2460
B-3001 Heverlee